UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

CONCEÇÃO DE UM DATA WAREHOUSE ESPÁCIO-TEMPORAL PARA ANÁLISE DE TRAJETÓRIAS HUMANAS

Vitor Hugo Fernandes Oliveira

PROJETO

MESTRADO EM INFORMÁTICA

2013

UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Informática

CONCEÇÃO DE UM DATA WAREHOUSE ESPÁCIO-TEMPORAL PARA ANÁLISE DE TRAJETÓRIAS HUMANAS

Vitor Hugo Fernandes Oliveira

PROJETO

MESTRADO EM INFORMÁTICA

Projeto orientado pela Professora Doutora Ana Paula Pereira Afonso
e pelo Professor Doutor André Osório e Cruz de Azerêdo Falcão

2013

Agradecimentos

Em primeiro lugar, quero agradecer à Professora Ana Paula Afonso pela confiança e disponibilidade demonstradas desde o primeiro dia em que me deu oportunidade de trabalhar neste projeto. Juntamente ainda com o Professor André Falcão, agradeço todo o apoio e motivação que me deram para continuar mesmo quando não sabia o rumo que deveria tomar no projeto. Um grande obrigado aos dois pela contribuição e ajuda na conclusão de mais uma etapa da minha vida académica.

O meu maior agradecimento vai para as duas pessoas que mais fazem por mim desde o dia em que nasci, os meus pais. Obrigado pelos sacrifícios que fizeram para me darem tudo e por fazerem de mim o homem que sou hoje. Obrigado pela paciência que tiveram quando as coisas correram mal e pelo o orgulho que demonstram quando alcanço os meus objetivos. Obrigado por tudo! Agradeço às minhas avós por todo o carinho demonstrado, ainda que uma delas já não esteja cá para me ver terminar esta etapa, e ainda à minha restante família pelo apoio que demonstrou quando foi necessário.

Um especial agradecimento à minha namorada Johanna que sempre me apoiou, aturou, e demonstrou uma enorme amizade e amor durante estes anos em que estivemos juntos. Obrigado por me fazeres encarar a vida com um sorriso todos os dias desde que estamos juntos. Agradeço ainda aos teus pais por tudo o que fazem por mim e pelo carinho que me dão.

Um grande obrigado por todos os momentos extra faculdade aos meus amigos de sempre, a Project Team. Por fim, agradeço a todos os meus amigos que criei durante estes anos de faculdade: Diogo Carvalho, João Félix, João Alves, André Francisco, João Alves, Bruno Pombeiro, Mafalda Gomes, Nélia Costa, entre outros. Obrigado por estes anos de vida académica. Um especial obrigado ao Rui Pires por todos estes anos de companheirismo, amizade e esforço, sem os quais teria sido muito mais difícil realizar esta etapa que finalmente estamos a concluir. Um obrigado ainda ao Miguel Garcia por toda a ajuda e disponibilidade durante este último ano de mestrado.

Para os meus pais.

Resumo

Com a evolução das tecnologias móveis à disposição dos utilizadores, tem ocorrido um aumento significativo do volume de dados produzidos a partir destes dispositivos. A disponibilização destas grandes quantidades de informação, por exemplo, sobre a localização de utilizadores móveis e respetivas trajetórias, potencia o conhecimento e o estudo sobre as atividades, preferências, padrões de comportamento e de mobilidade desses utilizadores no espaço e no tempo.

De modo a extrair informação útil e relevante é fundamental a conceção de métodos adequados para o tratamento, análise, descoberta de conhecimento e prospeção de dados. Contudo, os dados existentes sobre a mobilidade humana apresentam ainda redundâncias, incoerências, pouca informação semântica e ainda são escassas as soluções de descoberta de conhecimento e algoritmos de prospeção de dados especialmente concebidos para dados espácio-temporais.

Neste projeto é proposto um modelo de um *Data Warehouse* Espácio-Temporal de trajetórias humanas, assim como os processos necessários para o tratamento de dados e o seu enriquecimento com informação, tais como extração de pontos de estadia e um algoritmo para a descoberta de utilizadores semelhantes baseado em informação geográfica. Este modelo tem como finalidade criar as bases para a concretização de aplicações e algoritmos de deteção de comportamentos e atividades de utilizadores móveis, sendo testado num exemplo concreto, o conjunto de dados Geolife, para uma população de 182 utilizadores com cerca de 24 milhões pontos geolocalizados em trajetórias. Os resultados mostram que o sistema desenvolvido permite níveis de análise de grande complexidade, possibilitando simultaneamente uma grande flexibilidade para processamento analítico, apresentando a sua utilidade para processos de negócio como planeamento urbano, análise de tráfego e análise de perfil de utilizadores.

Palavras-chave: *Data Warehouse*, Dados Espácio-Temporais, OLAP, Trajetórias, ETL.

Abstract

With the evolution of mobile technologies available to users, there has been an significant growth of the volume of data generated from these devices. The availability of these large quantities of information, for example, about the location of mobile users and their trajectories, enhances the knowledge and study on activities, preferences, behavior patterns and mobility of those users in both space and time.

In order to extract useful and relevant information is critical to designing appropriate methods for processing, analysis, knowledge discovery and data mining. However, the existing data on human mobility have still redundancies, inconsistencies, poor semantic information and are still scarce solutions of knowledge discovery and data mining algorithms specially designed for this type of spatio-temporal data.

This thesis propose a model of a Spatio-Temporal Data Warehouse of human trajectories, as well processes required for data processing and enrichment with semantic information, such as extraction of stay points and an algorithm for finding similar users based on geographic information. This model aims to lay the groundwork for the development of applications and algorithms for detection of behaviors and activities of mobile users, being tested in a concrete example, the data set Geolife for a population of 182 users with about 24 million points geolocated trajectories. The results show that the developed system allows analysis levels of complexity, while allowing great flexibility for analytical processing, showing its usefulness for business processes such as urban planning, traffic analysis and users profile analysis.

Keywords: *Data Warehouse*, Spatio-Temporal Data, OLAP, Trajectories, ETL.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este primeiro capítulo apresenta a motivação e os objetivos do trabalho desenvolvido para este projeto. São ainda apresentadas as contribuições do trabalho realizado, a metodologia e planeamento seguidos, tal como a estrutura do documento.

1.1 Motivação

Desde o início da década de 2000 tem-se observado uma rápida evolução das tecnologias móveis, desde os telemóveis até aos dispositivos GPS (*Global Positioning System*). Atualmente as características dos *smartphones* continuam com um desenvolvimento bastante acelerado, sendo até provável que se possa chegar a um ponto de estagnação em relação à inovação dos mesmos. Esta evolução envolve a integração de diversas tecnologias, e atualmente estes dispositivos são utilizados como câmara fotográfica, leitor de música, dispositivo de navegação na Internet, navegação GPS, entre outros.

Como resultado desta evolução tecnológica, a captura de grandes volumes de dados referentes a utilizadores móveis, entre outros, tornou-se tecnicamente e economicamente viável, existindo assim um número crescente de novas aplicações com o objetivo de organizar, gerir e extrair informação relevante desses dados, nomeadamente o estudo de comportamentos humanos através das suas trajetórias. Uma trajetória é um caminho que um dado objeto móvel executa num determinado espaço e período de tempo. Tipicamente uma trajetória espácio-temporal é representada por um conjunto ordenado de pontos $P_i = (x_i, y_i, t_i)$ em que x_i e y_i representam as coordenadas geográficas no tempo t_i e compõe uma trajetória $T = \{P_1, ..., P_n\}$ [13, 38]. De modo a extrair informação útil e relevante destes conjuntos de dados é fundamental a conceção de métodos adequados para o tratamento e análise desses dados de modo a permitir a descoberta de padrões de comportamento dos utilizadores, análise de perfil, planeamento urbano, controlo de tráfego, *marketing*, entre outros. Contudo, o volume de dados produzidos por movimentos humanos é ainda es-

cassamente utilizado para análise, descoberta de conhecimento e prospeção de dados. As principais razões para este facto prendem-se com dois aspetos. O primeiro, relaciona-se com o facto dos dados sobre trajetórias tipicamente apresentarem redundâncias, inconsistências e possuem pouca ou nenhuma informação semântica, assim como não existe uma estrutura abstrata de representação. Por outro lado, existem poucos algoritmos especialmente concebidos para prospeção de dados e exploração de padrões para este tipo de objetos móveis. Uma das aproximações para a resolução do primeiro problema apresentado é a construção de um *Data Warehouse* (DW) [26].

Um DW possibilita a análise de grandes volumes de dados, sendo os mesmos organizados e manipulados de acordo com os conceitos e operadores fornecidos por um modelo de dados multidimensional que apresenta os dados na forma de um cubo de dados [17]. Um cubo de dados permite que os dados sejam modelados e visualizados em múltiplas dimensões, representando cada dimensão uma perspetiva de negócio e é tipicamente implementado através de um esquema em estrela (*star schema*) [17], que tem como centro uma tabela de factos. Através de operações OLAP (*On-Line Analytical Processing*) [26] é possível explorar os dados contidos no DW, utilizando várias perspetivas e níveis de granularidade.

Apesar de já existir alguns trabalhos relacionados que abordam aspetos da problemática da análise e modelação de trajetórias humanas [4, 9, 29, 38], ainda não existe um modelo de dados consistente para análise de comportamento de utilizadores móveis no espaço e no tempo, integrado com informação semântica de forma a aumentar a expressividade do modelo e simplificar a sua compreensão e utilização através de um cubo de dados multidimensional. De facto, a existência de um modelo de dados permitirá criar as bases para a concretização de aplicações e algoritmos de deteção de comportamentos, atividades de utilizadores móveis, e ainda demonstrar a sua utilidade nas áreas mencionadas.

1.2 Objetivos

Este projeto tem como objetivo a modelação e concretização de um *Data Warehouse* para dados espácio-temporais referentes a trajetórias de humanos com o intuito de facilitar a extração de informação. É apresentado o processo ETL (*Extract, Transform, Load*) [25] que efetua o correto tratamento do conjunto de dados e produz novas informações. Será também dada ênfase à criação de um cubo de dados que permita realizar uma análise multidimensional e hierárquica dos dados no DW, de forma a ser possível efetuar análises através de operadores OLAP.

Este projeto enquadra-se no projeto SInteliGIS (*Services for Intelligent Geographical Information Systems*) [2, 3], que propõe um programa de investigação na área da gestão

de informação georreferenciada, focando-se particularmente em problemas relacionados com a extração e a recuperação desta informação.

De seguida são apresentados os objetivos específicos propostos por este projeto, tal como uma breve descrição da abordagem aos mesmos:

1. **Criação de um modelo de um Data Warehouse Espácio-Temporal:** será proposto um modelo que permita a organização, gestão e extração de dados relativos a movimentos humanos, isto é, trajetórias humanas. O DW será modelado de acordo com os requisitos que os dados desta natureza implicam, tal como os requisitos apresentados pela literatura existente [17, 26, 45]. O modelo tem como principais desafios apresentar uma estrutura simples e coerente, possibilitando assim a fácil compreensão dos dados que armazena, tal como permitir a fácil integração com técnicas de análise OLAP.

2. **Modelação de um processo ETL:** o processo que será proposto deve permitir efetuar o tratamento de grandes volumes de dados referentes a movimentos de utilizadores móveis, sendo o seu principal objetivo a conceção de métodos referentes ao enriquecimento dos dados com informação semântica. Os métodos propostos são os seguintes:

   2.1. **Caracterização semântica de trajetórias:** este método deverá permitir a criação de informação semântica com base na trajetórias dos utilizadores móveis, tais como, velocidade e distância percorrida.

   2.2. **Agrupamento de localizações:** este método deverá permitir efetuar o agrupamento de registos geográficos pertencentes às trajetórias de utilizadores móveis, ou seja, a criação de grupos de registos que estejam próximos geograficamente.

   2.3. **Extração de pontos de estadia:** este método deverá permitir a extração de pontos de estadia das trajetórias dos utilizadores, ou seja, o método deverá interpretar uma trajetória e extrair os registos nos quais um utilizador permaneceu num local durante um certo período de tempo.

   2.4. **Descoberta de utilizadores semelhantes:** este método deverá permitir agrupar utilizadores com base nos seus pontos de estadia e localizações frequentadas, isto é, com base apenas em informações geográficas sobre as trajetórias dos utilizadores móveis, criar grupos de utilizadores semelhantes.

3. **Criação de um cubo de dados:** a aproximação de criar um cubo de dados é a mais apropriada para efetuar análises OLAP num DW. Desta forma, este deve permitir efetuar análises que permitam extrair informação relevante, tal como possibilitar a utilização de hierarquias e técnicas OLAP, tais como *drill-down*, *roll-up* e *pivot*.

1.3 Contribuições

Com base nos objetivos anteriormente mencionados, este projeto apresenta as seguintes contribuições:

- Criação de um modelo de dados consistente, integrado com informação semântica que permita criar as bases para a concretização de aplicações e algoritmos de deteção de comportamentos, atividades de utilizadores móveis, e ainda que demonstre a sua utilidade nas áreas de planeamento urbano, controlo de tráfego, análise de perfil, *marketing*, entre outras relacionadas.
- Aplicação do modelo a um conjunto de dados concreto que demonstra que o sistema desenvolvido permite níveis de análise de elevada complexidade, permitindo simultaneamente uma grande flexibilidade para processamento analítico. Este objetivo é concretizado através de um cubo de dados multidimensional resultante do modelo que será proposto.
- Avaliação e análise de resultados através do estudo de casos concretos que se adaptam às áreas nas quais o modelo demonstra a sua utilidade, tal como, a demonstração da integração do modelo com ferramentas de visualização de dados referentes à movimentação de utilizadores móveis.

Para além destas contribuições, foi ainda produzida uma publicação referente ao principal objetivo deste projeto. O artigo intitulado, Conceção de um Data Warehouse Espácio-Temporal para Análise de Trajetórias Humanas, na conferência INForum'13: Simpósio de Informática, que decorreu no mês de Setembro em Évora [30], tem como tema a conceção do modelo de dados do *data warehouse* espácio-temporal, explorando a modelação dimensional, tal como as técnicas de enriquecimento semântico que tinham sido desenvolvidas até à a data da realização do artigo.

1.4 Metodologia e Planeamento

De acordo com os objetivos apresentados, neste projeto foi utilizada uma metodologia de trabalho iterativa em que os diversos componentes do projeto estiveram em constante evolução e aperfeiçoamento, permitindo assim chegar à solução final desejada para este projeto. Apresenta-se, em seguida, as principais etapas da metodologia utilizada:

- A primeira etapa refere-se ao estudo do estado da arte (ver Capítulo 2), em técnicas de análise e exploração de trajetórias humanas, e mais concretamente em *data warehouses* espácio-temporais aplicados a trajetórias.

- Na segunda etapa é feito um levantamento dos requisitos e limitações do trabalho já existente na literatura, procurando-se apresentar soluções que cumpram os requisitos e abordagens que minimizem ou resolvam as limitações existentes (apresentado no Capítulo 3).

- Após as duas etapas anteriores, é feita a concretização da solução proposta na etapa anterior, sendo a sua construção efetuada em diversas iterações, em que a solução passará por um período de concretização e avaliação contínua até atingir os objetivos propostos (apresentado no Capítulo 4).

Tendo em conta a metodologia apresentada foi proposto e executado o planeamento do projeto (ver Anexo A) que envolveu as seguintes tarefas principais:

- A primeira tarefa, de duração aproximada de 1 mês, consistiu em uma primeira fase de análise do problema realizando-se de seguida uma pesquisa e análise do estado da arte do âmbito do projeto. Este estudo focou-se principalmente na pesquisa de abordagens existentes na literatura aos *data warehouses* espácio-temporais, sendo numa segunda fase estudada a aplicação destes a trajetórias de objetos móveis. No âmbito das trajetórias foram ainda explorados os trabalhos existentes referentes a trajetórias semânticas.

- A segunda tarefa, de duração aproximada de 2 meses, consistiu no aprofundamento dos requisitos de modelos de *data warehouses* espácio-temporais de trajetórias humanas, sendo realizada em várias etapas a modelação dimensional do modelo proposto neste projeto. Nesta segunda fase abordou-se também a questão de trajetórias semânticas, em que foram estudadas as características das mesmas e efetuada a modelação de processos que foram de encontro às necessidades que o âmbito do modelo apresentou.

- A terceira tarefa, de duração aproximada de 6 meses, consistiu na concretização do modelo que foi concebido na tarefa anterior. Este modelo foi aplicado a um conjunto de dados específico, que revelou novos requisitos e o que criou a necessidade de efetuar diversas iterações na construção. O *data warehouse* foi concretizado através de um processo ETL, no qual foram aplicados os processos desenvolvidos relativos às trajetórias semânticas. Após esta etapa, efetuou-se a avaliação experimental da concretização do modelo, através da implementação de um cubo de dados multidimensional e realizadas explorações do ponto de vista analítico ao conjunto de dados utilizado. A parte final desta tarefa envolveu a recolha das principais conclusões da avaliação.

- A elaboração da publicação mencionada na secção anterior, teve como duração 1 mês.

- Por fim, com a duração de 1 mês, foi efetuada a elaboração do presente documento.

1.5 Estrutura do Documento

Após este primeiro capítulo, este documento está organizado da seguinte forma:

- Capítulo 2 - Conceitos e Trabalho Relacionado: são apresentados os conceitos que este projeto aborda, tal como a literatura existente sobre os trabalhos relacionados com a concretização de DWs espácio-temporais aplicados a trajetórias humanas e outros objetos móveis.
- Capítulo 3 - *Data Warehouse* Espácio-Temporal: é apresentado o trabalho desenvolvido para a conceção do modelo para um *Data Warehouse* espácio-temporal. É detalhada a criação do modelo através das linhas gerais para a conceção de um DW. Serão ainda apresentadas as bases teóricas dos métodos de criação de informação semântica para trajetórias.
- Capítulo 4 - Validação do Modelo Proposto: é apresentada a concretização do modelo do DW espácio-temporal, tal como os métodos integrados no processo ETL do modelo apresentado. Esta concretização terá como base o conjunto de dados Geolife [41] da *Microsoft Research Asia* [1]. É efetuada a experimentação do modelo através de diversas interrogações analíticas que se integram nas áreas propostas, tal como a demonstração da integração do modelo com aplicações de visualização de dados de referentes a trajetórias.
- Capítulo 5 - Conclusões e Trabalho Futuro: são apresentadas as principais contribuições do trabalho desenvolvido e discutidos os aspetos que permanecem em aberto. Finalmente, serão abordadas possíveis matérias de futuro desenvolvimento do trabalho realizado.

[1] http://research.microsoft.com/en-us/labs/asia/

Capítulo 2

Conceitos e Trabalho Relacionado

Neste capítulo são apresentados os conceitos fundamentais, nomeadamente o conceito de trajetórias, espaço geográfico, *data warehouse* e por fim *data warehouse* espácio-temporal. Na segunda secção, é apresentado o estado da arte na área dos modelos para *data warehouses* espácio-temporais. Por fim, na terceira secção será feita uma discussão comparativa dos trabalhos apresentados com base em diversos critérios.

2.1 Conceitos

2.1.1 Trajetória

Por definição, uma trajetória consiste num caminho que um dado objeto móvel executa num determinado espaço e período de tempo [17]. Dado que um movimento nunca é feito instantaneamente, o fator tempo está interligado com o conceito de trajetória. Considerando que uma posição geográfica é dada por um x e y (tipicamente coordenadas geográficas como latitude e longitude) que variam no espaço temporal entre t_1 e t_i, podemos afirmar que um ponto de uma trajetória pode ser representado por P_i = (x_i, y_i, t_i) [13, 43]. Portanto, uma trajetória que varie no espaço e no tempo pode ser representada por $T = \{P_0, ..., P_i\}$ (ver Figura 2.1).

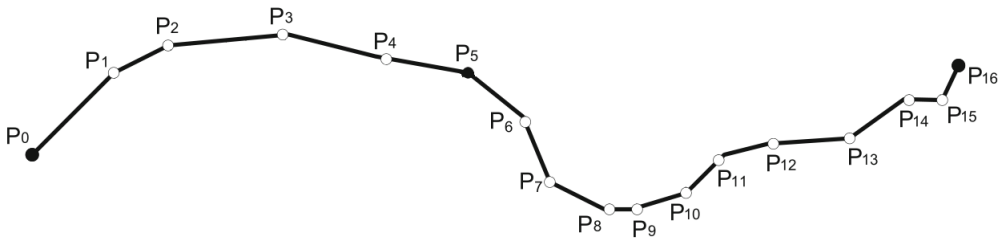


Figura 2.1: Exemplo de uma trajetória.

As sequências de pontos de uma trajetória podem resultar de diversas formas (ou uma combinação das mesmas) de observar movimentos e de recolha de dados de trajetórias:

- Registo baseado no tempo: os pontos são gravados regularmente em períodos específicos de tempo (por exemplo, de 5 em 5 segundos);
- Registo baseado no espaço: os pontos são gravados sempre que uma posição difere da posição anterior;
- Registo baseado na localização: os pontos são gravados sempre que o objeto móvel se encontra perto de uma localização especifica;
- Registo baseado em eventos: os pontos são gravados quando certos eventos ocorrem (por exemplo, fazer uma chamada).

Inconsistências nas Trajetórias

Um dos grandes problemas em relação a grandes quantidades de dados sobre movimentação de objetos deve-se tipicamente às inconsistências e redundâncias que a captura destes dados apresentam. Apesar da tecnologia de GPS ter evoluído bastante, esta continua a não apresentar um grau de precisão aceitável quando se trata de apresentar uma eficácia aceitável no registo da movimentação de um objeto. Este facto está relacionado com diversos fatores, tais como, interferências na captura da localização (por exemplo, prédios, túneis), condições meteorológicas, posicionamento dos satélites [33].

Estas condicionantes podem provocar erros de captura variáveis. Na existência de algum ruído o erro pode ser pequeno (por exemplo, aproximadamente 1 metro) mas nas grandes cidades onde existem grandes estruturas, ou até em desfiladeiros, a margem de erro pode-se multiplicar várias vezes, graças à existência de um maior ruído, causando assim uma menor precisão na captura. Outro problema associado, é a inexistência de sinal GPS dentro de edifícios. Na Figura 2.2 estes registos anormais (*outliers*) são facilmente verificados pois tipicamente são pontos que estão completamente distanciados do seguimento dos restantes pontos da trajetória.

Adicionalmente, existe ainda um problema relacionado com a forma de registo das trajetórias. Tal como referido anteriormente, uma trajetória pode ser registada de diversas formas, estando ainda a representação das mesmas condicionada pelo tipo e modelo do dispositivo de captura da movimentação (por exemplo, telemóvel, relógio de GPS, entre outros). O facto de não existir uma norma que defina uma estrutura de representação uniforme, condiciona a análise de trajetórias, e consequentemente exige um considerável investimento na análise e transformação dos dados antes de serem armazenados.

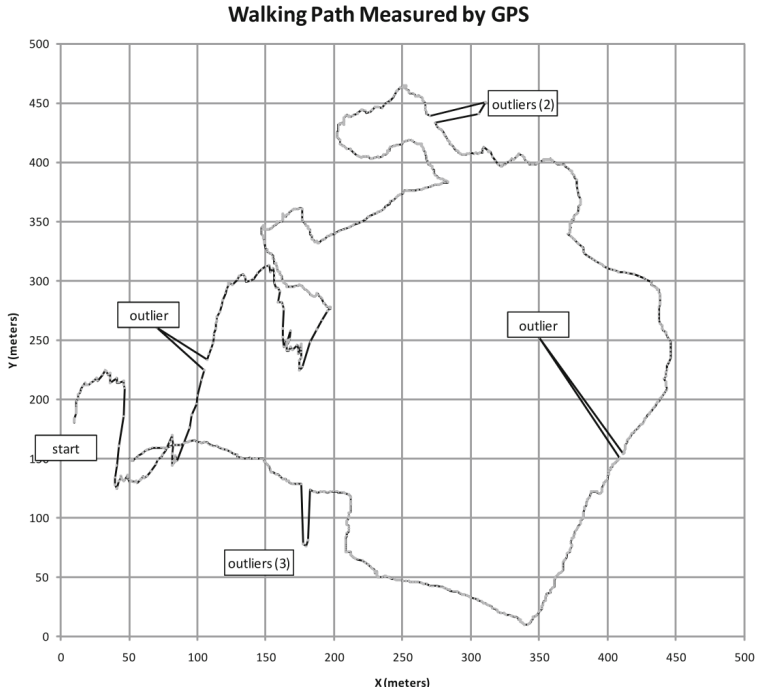


Figura 2.2: Exemplo de trajetória com registos anormais (*outliers*) [45].

Os problemas de representação acrescidos com os erros de captura, criam diversos problemas, nomeadamente ao investimento na conceção de métodos para análise e tratamento de dados, algoritmos de prospeção de dados, aplicações para descoberta de conhecimento, entre outros. Isto tem tido algum reflexo na escassa existência de métodos específicos para análise com trajetórias, e no pouco uso deste tipo de dados pelo nicho comercial/industrial que teria interesse nos mesmos.

Trajetórias Semânticas

Como já mencionado, um ponto de uma trajetória é caracterizado pelo seu registo espacial (posição no espaço) e temporal (posição na escala de tempo). No entanto, existem outras características que podem ser enumeradas, tais como, a direção do objeto, a sua velocidade momentânea, mudança de direção, aceleração, entre outros. Por sua vez, uma trajetória tem também diversas características associadas tais como a sua forma geométrica no espaço, a distância total percorrida no espaço, a duração temporal, as características de velocidade (média, mínimos, máximos) e seu comportamento (períodos de aceleração, desaceleração, ordem dos mesmos), e comportamento das direções (mudanças de direção e suas características) [8, 17].

As caraterísticas referidas acrescentam valor semântico aos registos, no entanto, existe ainda informação que pode ser extraída das trajetórias, tal como os pontos de estadia do objeto móvel que ocorreram na trajetória. Na Figura 2.3 podemos observar a extração de

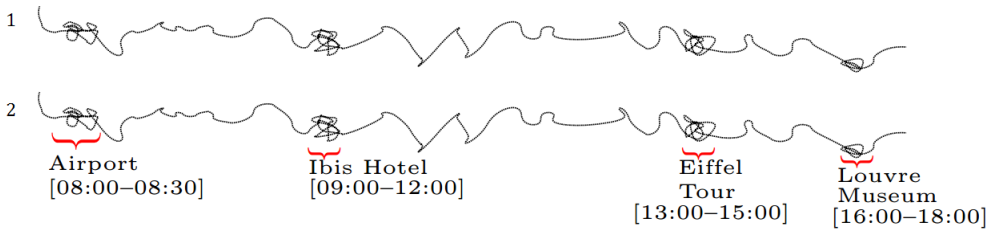


Figura 2.3: Exemplo de (1) trajetória abstrata (2) trajetória semântica [5].

informação de uma trajetória, observando em (1) uma trajetória sem qualquer informação, e em (2) uma trajetória com a informação dos pontos de estadia da trajetória, tal como do período temporal de estadia nesses pontos.

Um ponto de estadia, como pode ser observado na Figura 2.4, corresponde a duas situações que podem ocorrer numa trajetória: na primeira situação (Ponto de Estadia 1) o utilizador permanece num determinado local um certo limite de tempo (por exemplo, o utilizador entra num edifício perdendo assim o sinal de GPS, sendo este readquirido quando sai). Na segunda situação (Ponto de Estadia 2) o utilizador permanece numa determinada região por um certo limite de tempo, obtendo-se assim diversos pontos GPS nessa região (por exemplo, passear num pequeno jardim ou estar na paragem do autocarro).

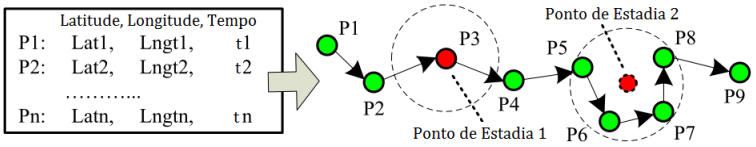


Figura 2.4: Representação de pontos de estadia.

A identificação exata de um ponto de estadia é um processo complexo, uma vez que identificar o ponto exato a que este realmente corresponde pode não ter a eficácia pretendida derivado a diversos fatores, tais como, erro de captura através do dispositivo de GPS, grande distribuição de pontos de interesse que existem nas grandes cidades. Estas situações são facilmente verificáveis na Figura 2.5. Zheng *et al.* [45] apresenta uma solução para este problema, mas apenas conseguiram calcular o ponto de interesse visitado por grupos de utilizadores, não conseguindo aproximar o ponto exato que foi visitado por um utilizador individual.

Outra informação semântica que é relevante na análise de trajetórias, é o tipo de movimento efetuado pelo objeto móvel. Por exemplo, uma pessoa durante o trajeto de casa para o trabalho, pode ter vários tipos de movimentos, tais como, andar a pé, autocarro e comboio. Estas informações são bastante relevantes para analisar o comportamento de um utilizador ou grupos de utilizadores durante as suas rotinas diárias, nomeadamente, entender o dinamismo do tipo de movimentação relacionado com as características já

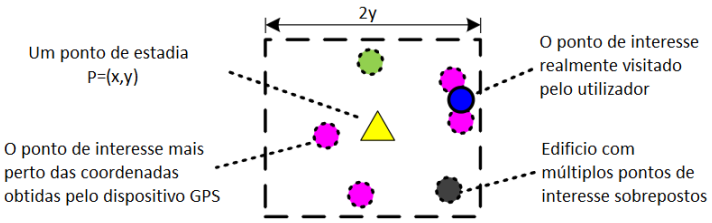


Figura 2.5: Erro associado à captura de pontos de estadia.

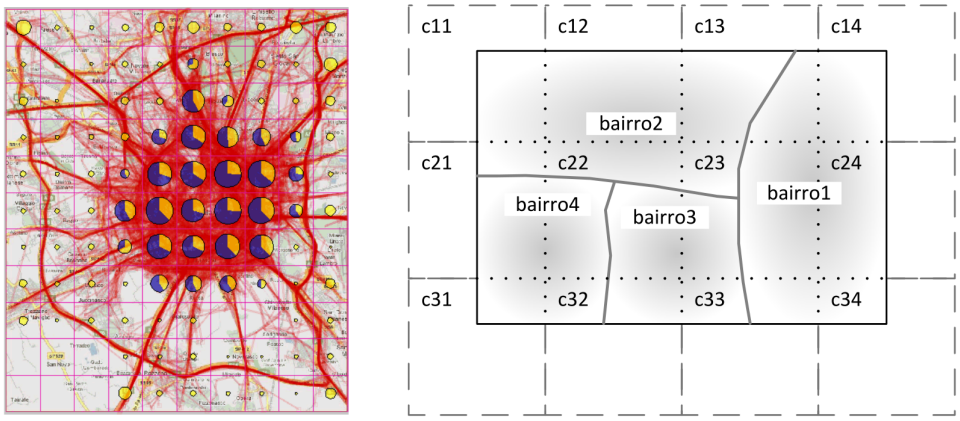
mencionadas (por exemplo, velocidade, aceleração). Zheng *et al.* [42] criaram uma abordagem para tentar inferir o tipo de movimentação numa trajetória, porém os resultados não são satisfatórios quanto à precisão do real tipo de movimento de um utilizador, o que invalida a sua utilização do ponto de vista analítico.

Ainda do ponto de vista analítico, é fundamental a comparação da movimentação de objetos (por exemplo, pessoas diferentes), surgindo então a necessidade de comparação de trajetórias, quer seja em diferentes registos temporais (por exemplo, trajetórias de uma pessoa em diferentes dias) quer seja em diferentes partes da mesma trajetória (por exemplo, trajetórias de uma pessoa de casa para o trabalho e vice-versa). Para estas comparações, o interesse recai sobre as já mencionadas características de trajetórias, e nas relações espaciais e temporais como a sua co-existência no espaço (têm as mesmas posições ou algumas posições em comum), co-existência no tempo (trajetórias feitas durante o mesmo período de tempo ou com alguns períodos semelhantes), e co-incidência no tempo e/ou no espaço (mesmas posições obtidas no mesmo tempo ou com um certo tempo de diferença) [17].

2.1.2 Espaço Geográfico

A formalização da representação do espaço geográfico quando se lida com dados espaciais é essencial. Uma abordagem possível é associar informação semântica para cada registo de uma trajetória, como por exemplo a rua e o número de um dado bairro. Porém, nem sempre é possível obter estas informações semânticas devido ao facto de não existir autorização para o acesso das mesmas, ou devido a problemas de captura nos dados.

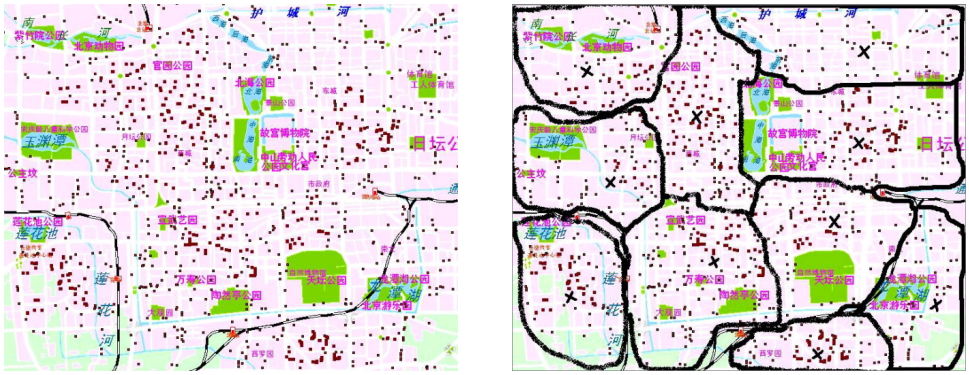
Giannoti *et al.* [17] refere que a representação geográfica de dados espaciais permite a divisão dos diversos registos por setores pertencentes à região abrangida. Esses setores podem ser distribuídos pela divisão de uma região numa grelha regular (Figura 2.6(a)), o que facilita a comparação entre setores visto que os mesmos possuem dimensões iguais. Neste tipo de representação surge tipicamente o problema do relacionamento parcialmente contido [23]. Supondo que uma região está dividida por células, este problema prende-se com o facto de uma célula poder conter mais do que um bairro, ou por outro lado, não conter a totalidade do mesmo, como pode ser observado na Figura 2.6(b).



(a) (b)

Figura 2.6: (a) Representação geográfica por grelha regular [35]. Os gráficos circulares representam a origem (em amarelo) e o destino (em azul) das trajetórias num dado espaço temporal. (b) Problema do relacionamento parcialmente contido [4].

Outra alternativa é o tratamento dos dados espaciais através de técnicas de agrupamento [7, 24]. Estas técnicas têm como propósito, em relação a dados espaciais, agrupar objetos espaciais similares em grupos ou classes (Figura 2.7), podendo ser usadas para a identificação de áreas de utilização similar num conjunto de dados espaciais e criação de conjuntos de regiões com condições meteorológicas idênticas, entre outros [20]. Neste trabalho será utilizada uma técnica de agrupamento hierárquico para a criação de grupos de localizações com base no conjunto de dados geográficos a ser utilizado no DW.



(a) (b)

Figura 2.7: (a) Registos espaciais (b) agrupados através de técnicas de agrupamento [14].

Técnicas de cálculo de distâncias

Na secção anterior foi referido que certas características de trajetórias poderiam ser extraídas através da análise e tratamento dos seus pontos. Dado que estamos a lidar com dados espaciais, informações como a aceleração e velocidade são calculadas com base

nas distâncias entre registos espaciais. Neste relatório serão abordadas duas técnicas para o cálculo destas distâncias, a fórmula de Haversine e a distância euclidiana.

Fórmula de Haversine Observando a Fórmula 2.1, tem-se que ϕ_i, λ_i correspondem às coordenadas (latitude, longitude) iniciais, ϕ_f, λ_f as respetivas coordenadas finais, e $\Delta\phi$, $\Delta\lambda$ a diferença entre as mesmas. A distância angular entre os pontos é representada por $\Delta\hat{\sigma}$, sendo necessário o cálculo de $r * \Delta\hat{\sigma}$ para a transformação da distância em quilómetros, em que que r representa o raio da Terra (aproximadamente 6 378.1km[1]) [37].

$$\Delta\hat{\sigma} = 2\arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\phi i\cos\phi f\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{2.1}$$

Distância Euclidiana A distância euclidiana entre dois pontos corresponde ao comprimento do caminho que os une. Pode ser calculada diretamente usando a fórmula pitagórica (Fórmula 2.2).

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \tag{2.2}$$

2.1.3 Data Warehouse

O conceito de DW [26] surge no seguimento da necessidade das grandes empresas conseguirem gerir grandes volumes de dados existentes, e consequentemente simplificar o processo de análise dos mesmos de forma a auxiliar no processo de tomada de decisão dos executivos das empresas. Surge então na década de 80 os *Executive Information Systems* (EIS) que permitem a agregação e atualização automática de múltiplas fontes de dados do mesmo repositório, dando assim origem aos DWs atuais. Existem diversas definições para o conceito de *data warehouse*, mas este projeto irá adotar a simples definição fornecida por Kimball *et al.*:

> *Um data warehouse é uma cópia de uma transação especificamente estruturada para pesquisa e análise [26].*

Tipicamente, um DW é elaborado com vista a aglomerar várias fontes de dados num só repositório, simplificando assim o seu acesso. Estes dados devem ser coerentes, passando por um longo processo de tratamento até serem armazenados, pois estes não ficam suscetíveis a alterações. Essas alterações podem ser feitas com base em registos

[1] http://nssdc.gsfc.nasa.gov/planetary/factsheet/earthfact.html

históricos, sendo esta tarefa um grande desafio em matéria de *data warehousing*. O DW deve ser elaborado de forma a possibilitar um fácil entendimento por parte dos executivos/decisores da empresa em questão, usando termos conhecidos pelos mesmos. Todos os dados e organização dos mesmos devem estar estruturados para melhorar, automatizar e tornar mais rápido os processos de tomada de decisão.

Num DW os dados são organizados e manipulados de acordo com os conceitos e operadores fornecidos por um modelo de dados multidimensional que os apresenta na forma de um cubo de dados [17]. Estes cubos são pré-calculados, o que resulta em respostas com um melhor desempenho, facto que é crítico para o uso executivo de sistemas OLAP [26]. Cada face do cubo representa uma dimensão que tem como objetivo representar uma entidade independente, sendo caracterizadas por ter muitas colunas e atributos. Cada ponto do cubo representa uma medida que está contida na tabela de factos, sendo esta a tabela primária num modelo dimensional onde as medidas numéricas (atributos que servem para avaliar o negócio) são guardadas [26] agregando também as chaves estrangeiras das tabelas de dimensão de forma a expressar um certo facto, sendo o seu significado designado por granularidade, que determina o nível máximo de detalhe [22].

A tabela de factos e as dimensões são tipicamente representados através de um esquema em estrela (Figura 2.8), em que a tabela de factos fica colocada no centro, estando rodeada pelas dimensões que a constituem.

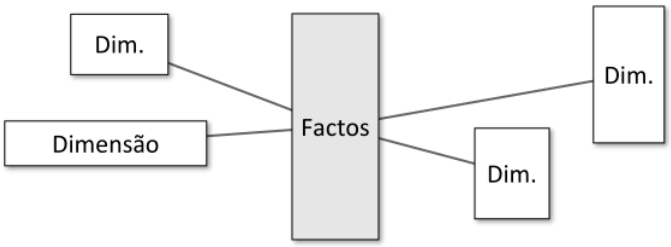


Figura 2.8: Representação de um esquema em estrela [16].

Através de operações OLAP é possível explorar os dados contidos no DW, utilizando várias perspetivas e níveis de granularidade. As operações OLAP [10] auxiliam a análise dos dados, sendo essenciais na manipulação das hierarquias das dimensões. As operações típicas nestes sistemas incluem a operação *Pivot* que permite a escolha da vista com dimensões pertinentes, *Roll-up* que permite aumentar o nível de agregação de resultados (obtendo assim resultados com menor detalhe), *Drill-down* que é o inverso da operação anterior (permitindo assim resultados com maior detalhe), *Slice* que permite restringir uma dimensão da análise, ou seja, permite selecionar as dimensões que fazem parte de uma análise, e por fim, a operação *Dice* que permite restringir valores de uma ou mais dimensões da análise.

Ciclo de Vida

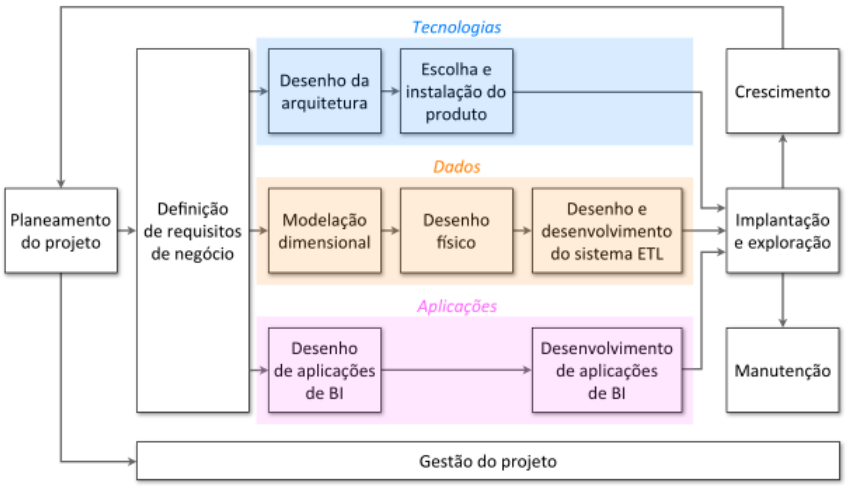


Figura 2.9: Diagrama do ciclo de vida de um *data warehouse* [16].

O ciclo de vida de um DW (Figura 2.9) além das típicas fases de planeamento e gestão do projeto, tem também a fase de definição de requisitos do negócio, que permite criar as bases para efetuar a modelação dimensional do mesmo. Este processo é essencial para a construção de um DW com um modelo de dados coerente e consistente, sendo tipicamente composto por quatro passos:

1. Listar prioridades para a construção do *data warehouse*: É um processo vital para o sucesso do DW. Se necessário define-se a matriz de exequibilidade/valor que salienta os processos com dados pesquisáveis, ou seja, pode-se observar os processos que realmente irão interessar.

2. Determinação do nível de detalhe da tabela de factos: Neste passo determina-se o nível de detalhe da tabela de factos, ou seja, define-se a sua granularidade. Este entende-se pelo significado de uma linha de tabela de factos.

3. Modelação das dimensões de negócio: Constrói-se a matriz de processos que relaciona as dimensões que foram definidas no passo anterior com os processos de negócio definidos no 1º passo. Após este passo, detalha-se as dimensões em relação aos seus atributos e ao seu tipo de dimensão.

4. Identificação das medidas numéricas de tabela de factos: Por último, é feita a identificação das medidas numéricas da tabela de factos. Também é concretizado o(s) esquema(s) em estrela que permite demonstrar a junção das tabelas de dimensões com as tabelas de factos.

Uma das fases mais importantes e que consome mais tempo (cerca de 70% [26]) na construção de um DW é o desenho e desenvolvimento do sistema ETL [25]. Este processo consiste na extração de dados de uma ou mais bases de dados, na transformação e limpeza dos mesmos, e no seu carregamento para o DW. É composto pelos seguintes passos:

- Extração (*Extraction*): Tem como objetivo a análise do domínio e das regras de integridade das colunas. É neste passo que se faz a deteção das alterações nos dados, a aplicação dos filtros e o ordenamento dos dados.
- Transformação (*Transformation*): Nesta fase faz-se a limpeza dos dados, o tratamento de exceções, a fusão de duplicados e a conformação de valores.
- Carregamento (*Load*): Esta é a parte final do processo, em que se faz o carregamento dos dados para o DW. Nesta fase é necessário manter as chaves substitutas (identifica univocamente cada linha da dimensão, não tendo qualquer ligação com os identificadores dos sistemas operacionais [26]), lidar com as dimensões de mudança lenta, preencher hierarquias e pré-calcular valores agregados.

Elementos Básicos

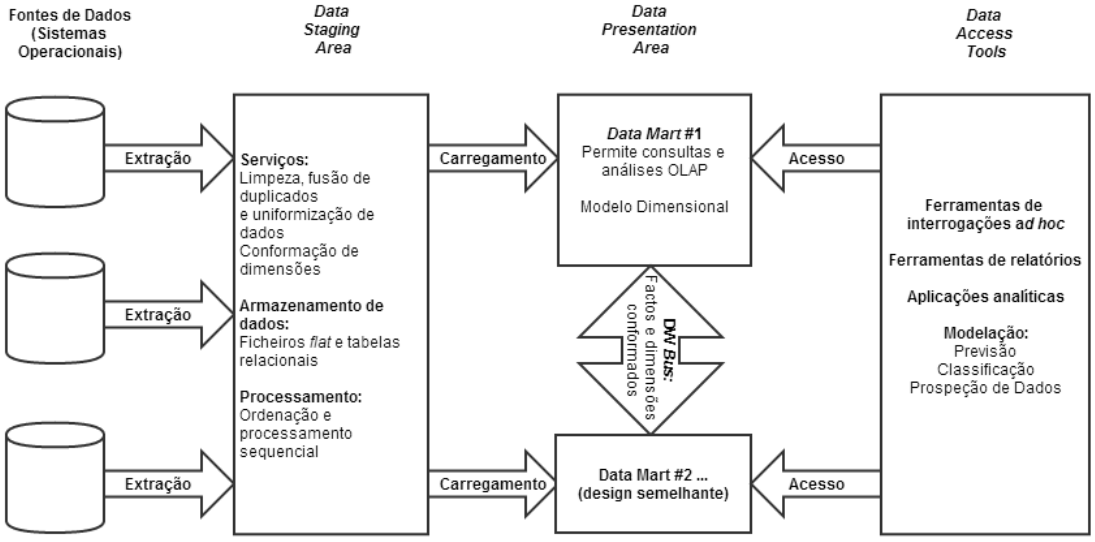


Figura 2.10: Elementos básicos de um *data warehouse* [26].

Na Figura 2.10 podemos observar os componentes que tipicamente constituem um DW. Um *data mart* é definido como um subconjunto de um DW, sendo orientado para um processo de negócio em especifico [26]. Os elementos básicos de um DW são então os seguintes:

- Fontes de dados (sistemas operacionais): Estes sistemas são as fontes de dados para o DW. São tipicamente externos ao DW, contendo o registo das transações do negócio mas sem manter um registo histórico [26].

- *Data Staging Area*: Esta é a área de trabalho que alberga os dados e consequentemente o processo ETL. É direcionada para pré-processar dados em bruto [16], sendo nesta área que são efetuadas operações típicas de limpeza dos dados, fusão de duplicados, combinação de dados e transformação, entre outras.

- *Data Presentation Area*: Área em que os dados estão organizados, armazenados, e disponibilizados para acesso direto dos utilizadores, executivos e aplicações analíticas. É nesta área que é feita a modelação dimensional do DW [26].

- *Data Access Tools*: Nesta área são albergadas todas as ferramentas que podem aceder à *data presentation area*, em particular aplicações analíticas e técnicas de prospeção de dados.

2.1.4 Data Warehouse Espácio-Temporal

Também definido como *Data Warehouse* de Trajetórias (DWTrs) [17], estes DW têm como motivação a transformação de informações sobre trajetórias em estado bruto em informação relevante e valiosa que possa ser utilizada para fins de tomada de decisão como por exemplo, em aplicações ubíquas, como serviços baseados em localização, controlo de tráfego, e planeamento urbano, entre outros [17]. Como já mencionado anteriormente, os dispositivos de GPS produzem um grande volume de dados em bruto, aumentando a complexidade dos dados a serem guardados em bases de dados de trajetórias, e simples interrogações feitas pelos utilizadores podem tornar-se complexas e computacionalmente dispendiosas.

Assim a abordagem de utilização de sistemas OLTP para extração de informação de trajetórias pode tornar-se num processo extremamente moroso e complexo. A solução para este problema tem sido a de aplicar os principios OLAP a estes sistemas, podendo assim aproveitar as vantagens destes e possibilitar a obtenção e análise de dados com significado para o utilizador do sistema.

Na Figura 2.11 podemos observar um exemplo de um DW espácio-temporal para trajetórias, verificando que as dimensões básicas representam o espaço, o tempo, o objeto móvel e o dispositivo de captura. A tabela de factos possui ainda medidas numéricas que caracterizam a trajetória.

Um DW temático, como um DW espácio-temporal, possui alguns requisitos aos quais deve obedecer durante o processo de modelação dimensional. De seguida serão então

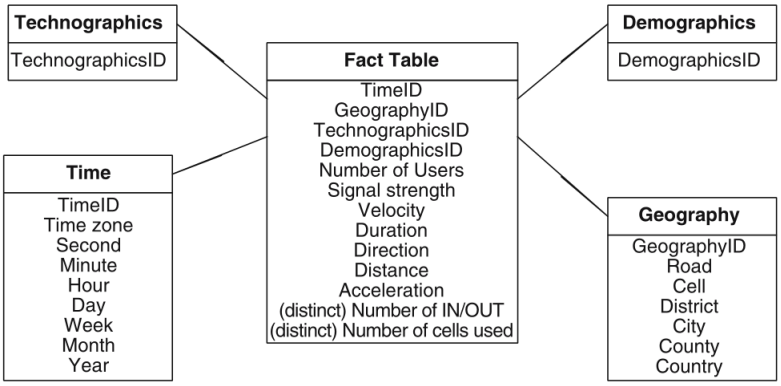


Figura 2.11: Exemplo de um *data warehouse* espácio-temporal [45].

analisados alguns pontos relevantes.

Dimensões espaciais, temporais e temáticas: para além das dimensões temporais (por exemplo, segundo, minuto, hora, dia, mês, ano) e espaciais (por exemplo, coordenadas, rua, cidade, pais), pode ser necessário existir dimensões temáticas que descrevam outros tipos de informação. Estas dimensões podem ser demográficas, nas quais são efetuadas a caracterização do utilizador ou objeto a quem o movimento pertence, ou ser tecnográficas em que se caracteriza o dispositivo de captura de movimento. Isto irá permitir caracterizar o movimento de forma qualitativa e não apenas quantitativa, aumentando assim a qualidade semântica da informação disponível no DW.

Medidas espaciais, temporais e temáticas: As trajetórias também podem possuir propriedades temáticas que são dependentes das suas características espaciais e temporais [17]. Assim, podem ser consideradas as seguintes características: (1) características numéricas, como a velocidade média da trajetória, a sua direção, a sua duração, a sua aceleração; (2) características espaciais, como a forma geométrica da trajetória; e (3) características temporais, como a temporização do movimento. O cálculo de algumas destas características pode ser computacionalmente dispendioso, podendo exigir pré-calculos durante o processo de análise das trajetórias.

Hierarquias: para possibilitar a utilização de técnicas OLAP, é necessária a criação de hierarquias. Com base nas dimensões referidas, existem diversas hierarquias que podem ser criadas, tais como temporais (por exemplo, Ano $>$ Mês $>$ Dia $>$ Hora $>$ Minuto $>$ Segundo) e espaciais (por exemplo, Pais $>$ Cidade $>$ Bairro $>$ Rua). Outra possibilidade, é a criação de grupos para representar diferentes níveis de abstração, como por exemplo, *género* = *feminino*, $idade = 25 - 35$, $estadocivil = solteira$ [17]. A criação de hierarquias e da correta definição dos atributos das dimensões, irá permitir uma fácil utilização dos operadores OLAP referidos na Secção 2.1.3. Por exemplo, na hierarquia atrás mencionada com o operador *roll-up* pode-se efetuar abstrações das análises como do nível $bairro$, para o nível $cidade$; com o operador *drill-down*, podemos detalhar uma

análise, como por exemplo, do nível $cidade$ para $bairro$; a operação de *slice* permite efetuar uma seleção sob uma dimensão, como por exemplo, $cidade = Lisboa$; por último, a operação *dice* permite mais que uma aplicação da operação slice, como por exemplo, $cidade = Lisboa$ e $ano = 2011$.

2.2 Modelos Espácio-Temporais

Os modelos para *data warehouses* espácio-temporais têm sido pouco explorados na literatura, mas existem algumas propostas que irão ser apresentadas de seguida.

Braz *et al.* [9] propõem um modelo de DWTrs, baseado num esquema em estrela com a granularidade da tabela de factos a representar uma célula de um dado espaço e tempo, sendo constituída por medidas numéricas que caracterizam o movimento da trajetória, tal como medidas de quantidade de trajetórias que estão presentes e que começam na célula, entre outras (Figura 2.12). As dimensões representam o espaço (*X_dimension* e *Y_dimension*) e o tempo (*Time_dimension*). Cada uma das dimensões possui hierarquias, em que cada nível é baseado em intervalos de espaço e de tempo (por exemplo, nível 0 > nível 1 > nível 2, como pode ser observado na Figura 2.13(b) da proposta de Orlando *et al.*), dependendo da dimensão. Estas dimensões não armazenam qualquer tipo de informação semântica, nomeadamente dados espaciais ou pontos de estadia das trajetórias, não permitindo a sua análise multidimensional.

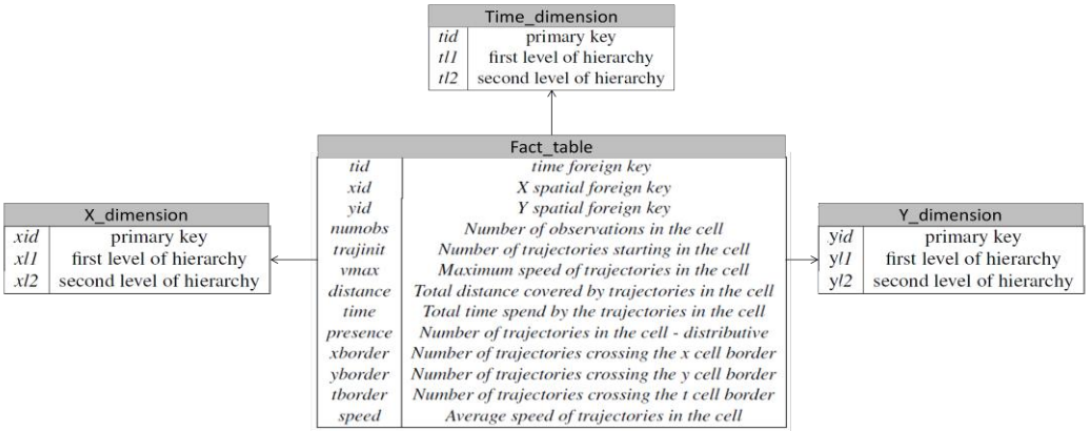


Figura 2.12: Esquema em estrela apresentado por Braz *et al.* [9].

Orlando *et al.* [31] apresentam um esquema em estrelade um DW (Figura 2.13(a)) com dimensões que representam o espaço geográfico (*dimX* e *dimY*), e uma dimensão temporal (*dimT*). À semelhança de Braz *et al.* [9], as dimensões possuem hierarquias baseadas em intervalos (Figura 2.13(b)). A tabela de factos apresenta apenas as medidas numéricas distância percorrida, velocidade a aceleração máxima (não presentes no

esquema apresentado), e medidas que registam as presenças das trajetórias no espaço geográfico.

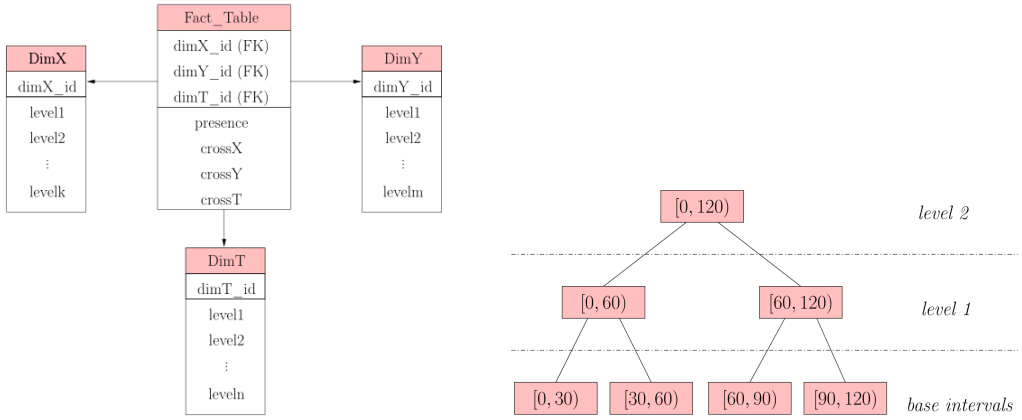


Figura 2.13: (a) Esquema em estrela de um DW espácio-temporal apresentado por Orlando *et al.* [31] (b) Hierarquização baseada em intervalos [31].

O modelo sugerido por Almeida *et al.* [4] foca-se essencialmente na gestão de tráfego urbano adotando o conceito de trajetórias semânticas. Este modelo foi o mais completo encontrado na literatura (Figura 2.14), possuindo dimensões que representam o espaço geográfico (*CelulaDim* e *Regiao*), a trajetória (*TrajDim* e *DirMovDim*, por exemplo), e o objeto móvel (*ObjMovDim*), tendo nas dimensões apropriadas as respetivas hierarquias. O modelo possui diversas informações semânticas, como por exemplo, pontos de estadia, informações do espaço geográfico, e do utilizador/objeto que realizou um movimento. O modelo tem a particularidade de possuir duas tabelas de factos: *Movimento Fato* que possui as medidas velocidade média, espaço percorrido, tempo decorrido, entre outras, e a tabela *ParadaFato* que caracteriza o ponto de estadia de forma sumária, através do tempo de parada e local de parada, entre outros. O modelo proposto permite ainda a análise multidimensional de trajetórias e contagem distinta, tendo também um módulo de SIG (Sistema de Informação Geográfica) integrado. Contudo, a análise dos dados a partir do modelo é feita através de interrogações OLTP, não sendo o modelo integrado num cubo de dados.

Em relação ao processo ETL de trajetórias, Marketos *et al.* [29] apresentam uma *framework* para a criação de um DW orientado a trajetórias (Figura 2.15) com as seguintes técnicas: (1) um método para a reconstrução de trajetórias durante o processo de carregamento de uma base de dados de movimentos espaciais; (2) métodos para a caracterização dos dados dos movimentos; e (3) a agregação de medidas para análises OLAP. O modelo do DW em que é aplicado o processo ETL proposto é similar ao modelo de Braz *et al.* [9], mas com algumas alterações que permitem análise de trajetórias (Figura 2.16): possui uma dimensão geográfica (*SPACE_DIM*) que embora não armazene dados espaciais

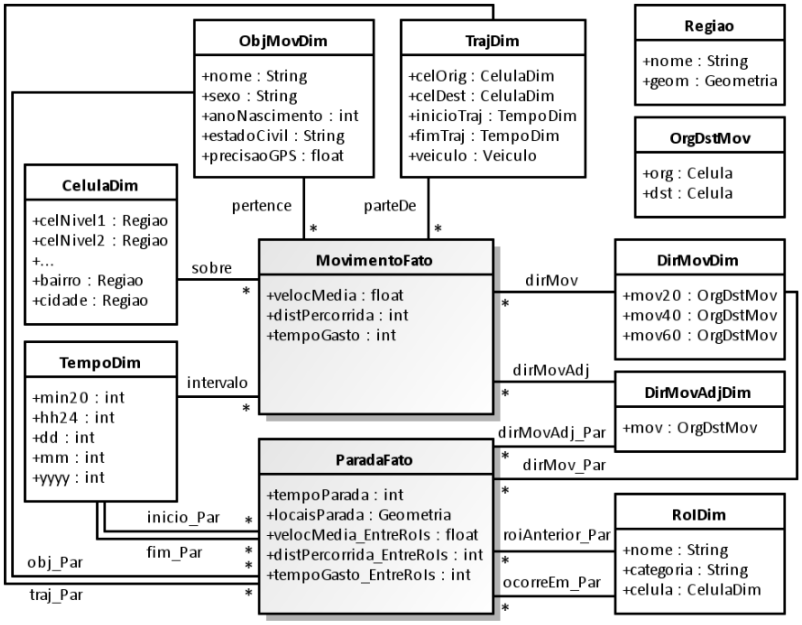


Figura 2.14: Esquema em estrela apresentado por Almeida *et al.* [4].

(o espaço é caracterizado por células), possuem informações geográficas dos mesmos e respetiva hierarquia; possui uma dimensão temporal (*TIME_DIM*) com a respetiva hierarquia associada; e possui uma dimensão para caracterizar o objeto móvel e o dispositivo de captura (*OBJECT_PROFILE_DIM*).

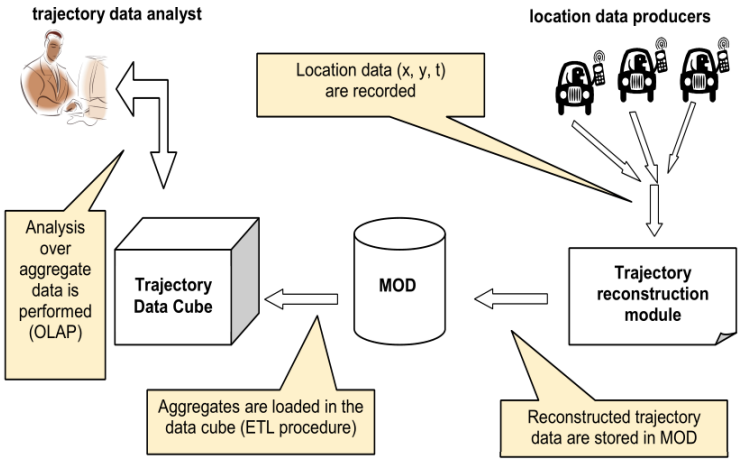


Figura 2.15: *Framework* apresentada por Marketos *et al.* [29].

Spaccapietra *et al.* [38] sugerem um modelo concetual para o enriquecimento semântico de trajetórias, em que estas possuem uma face geométrica com a sua sequência de registos, e uma face semântica, como pontos de estadia, e princípio e fim de uma trajetória. Este modelo para além de guardar dados espaciais, permite a contagem distinta de trajetórias e utilizadores/objetos (através de identificadores). Porém, não possui qualquer tipo de agregação e hierarquização de informação, possui o problema do relacionamento

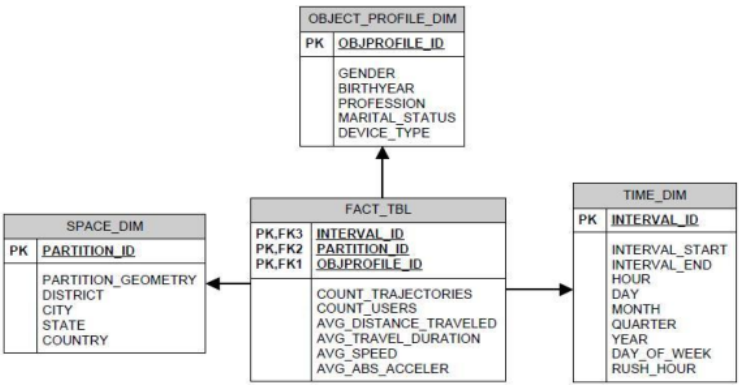


Figura 2.16: Esquema em estrela apresentado por Marketos *et al.* [29].

parcialmente contido, e o facto de ter um método que simplifica a trajetória (origem e destino do movimento, por exemplo) implica a perda de propriedades da trajetória.

Silva *et al.* [36] sugerem uma ferramenta ETL com o objetivo de diminuir os problemas existentes nesse processo da construção de um DW orientado a trajetórias. O modelo proposto (Figura 2.17) para aplicação da ferramenta, foca-se essencialmente no espaço geográfico: possui duas dimensões geográficas (*dimX* e *dimY*) que correspondem às coordenadas, longitude e latitude, respetivamente; e possui uma dimensão temporal (*dimT*). A hierarquização destas dimensões é feita por intervalos, tal como apresentam Braz *et. al.* [9]. A tabela de factos possui medidas que caracterizam numericamente as trajetórias, possuindo também a medida *presença* que representa a quantidade de trajetórias na célula analisada, tal como proposto por Braz *et. al.* [9]. A ferramenta desenvolvida (Figura 2.18) propõe que a hierarquização seja definida pelo utilizador, trabalhando com um formato definido de dados de entrada extraídos do GPS, efetua sempre um carregamento total dos dados (os dados antigos não são considerados) e possui as suas etapas definidas de acordo com o processo ETL. Tal como algumas das propostas anteriores, o modelo não possui bases para integração de informação semântica e as hierarquias presentes são baseadas em intervalos, podendo dificultar a sua interpretação.

2.3 Discussão

Após terem sido descritos os trabalhos mais relevantes existentes na literatura, pretende-se agora efetuar uma breve comparação entre eles, considerando as seguintes características:

1. Trajetórias semânticas: esta é uma das características fulcrais de um DW que seja orientado a trajetórias, implicando que o modelo proposto possua características designadas na Secção 2.1.1 (por exemplo, velocidade, aceleração, entre outras). Por exemplo, estas características são as que possibilitam que o modelo tenha utilidade

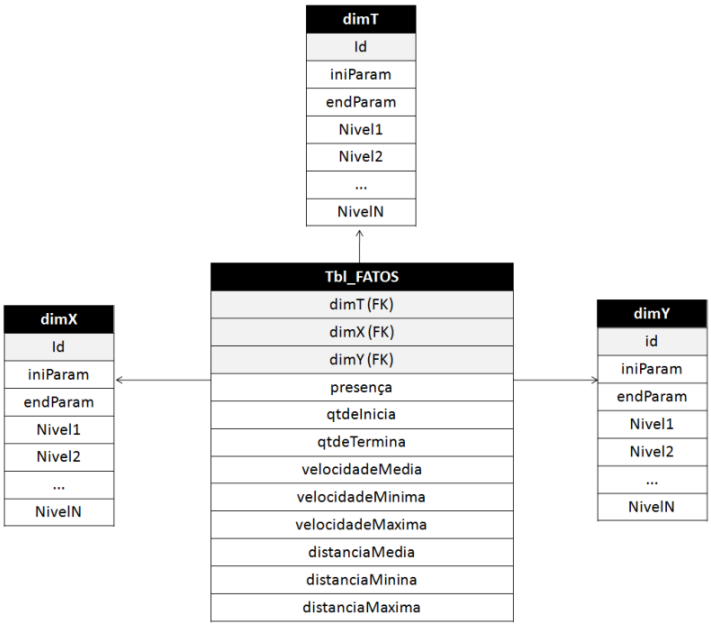


Figura 2.17: Esquema em estrela apresentado por Silva *et al.* [36].

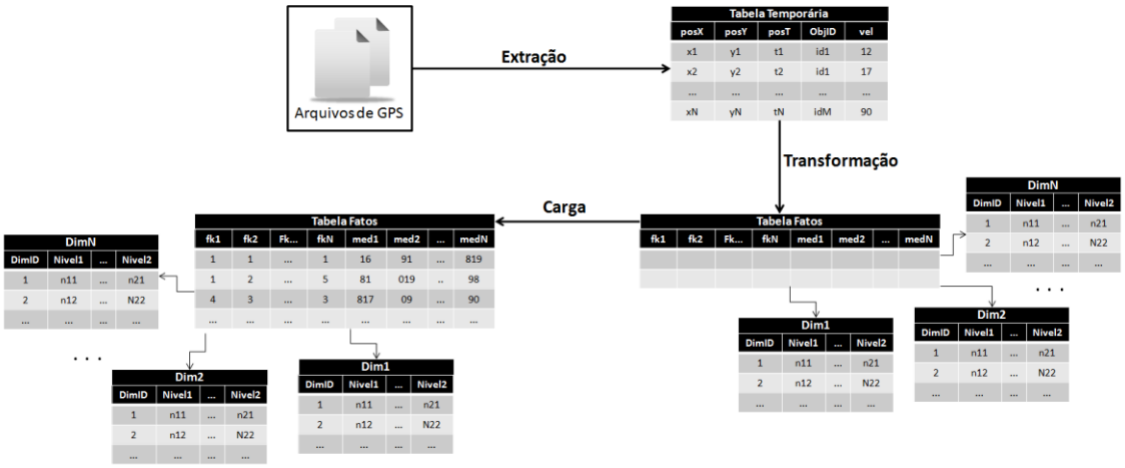


Figura 2.18: Etapas da ferramenta de ETL apresentadas por Silva *et al.* [36].

na gestão de tráfego urbano.

2. Hierarquização de informação: o modelo tem que possuir hierarquias nas suas dimensões, por forma a que seja possível efetuar análises OLAP com diferentes níveis de detalhe, tal como referido na Secção 2.1.4.

3. Modelo semântico: para além das características semânticas associadas às trajetórias, o modelo tem que possuir atributos nas suas dimensões. Por exemplo, no espaço geográfico (por exemplo, rua e cidade) as informações temporais (por exemplo, se é ou não um dia útil), entre outros.

4. Caracterização do objeto móvel: o objeto ou utilizador móvel tem de ser carac-

terizado no modelo, possuindo pelo menos um identificador numérico. Porém, é importante referir que a não caracterização do objeto móvel é um aspeto negativo, pois na literatura existente um dos aspetos debatidos relaciona-se com a privacidade dos dados [11, 17, 18].

5. Contagem distinta: o modelo proposto não deverá possuir o problema da contagem distinta (Figura 2.19), também conhecido por *distinct count problem* [9]. Este problema ocorre quando na agregação de registos de trajetórias existem erros na contagem distinta de objetos móveis. Por exemplo, um objeto pode ser contabilizado múltiplas vezes na mesma interrogação, pelo facto de o objeto possuir identificador. Modelos como o de Braz *et al.* e Orlando *et al.* diminuem este problema através das medidas de contagem de presença de trajetórias nas células do espaço geográfico.

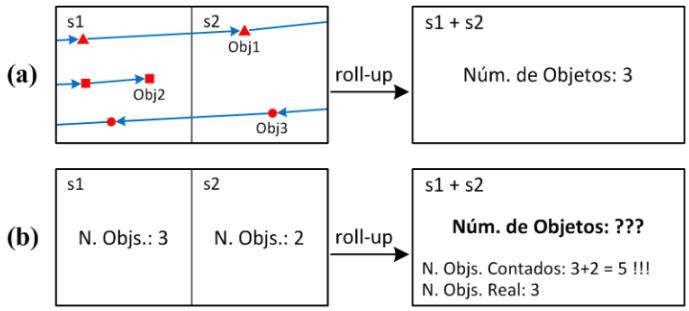


Figura 2.19: Agregação de dados de uma trajetória, em que (a) os dados não estão agregados (b) os dados estão agregados, ocorrendo o problema de contagem distinta [4].

6. Reconstrução de trajetórias: o modelo deve apresentar no seu processo ETL o método de reconstrução de trajetórias. Este método deverá extrair as trajetórias e pontos de estadia que fazem parte da mesma, a partir de dados em bruto de movimentação de objetos móveis.

7. Relacionamento parcialmente contido: o modelo deve resolver o problema do relacionamento parcialmente contido descrito na Secção 2.1.2.

8. Integração com SIG: o modelo sugerido deve possuir a capacidade de interligação com ferramentas de visualização de registos no espaço geográfico.

9. Distinção de pontos de estadia: o modelo deve apresentar no seu processo ETL o método para a extração de pontos de estadia das trajetórias dos objetos móveis, tal como apresentada na Secção 2.1.1.

A Tabela 2.1 apresenta uma comparação dos diversos modelos apresentados através da análise destes. A última coluna refere-se ao modelo proposto neste projeto e as suas características.

	Trabalhos						
Características	Braz *et al.*	Orlando *et al.*	Almeida *et al.*	Marketos *et al.*	Spaccapietra *et al.*	Silva *et al.*	Modelo Proposto
Trajetórias semânticas	✓	✓	✓	✓	✓	✓	✓
Hierarquização de informação	✓	✓	✓	✓		✓	✓
Modelo semântico			✓		✓		✓
Caracterização do objeto móvel			✓				✓
Contagem distinta	✓*	✓*	✓	✓*	✓	✓*	✓
Reconstrução de trajetórias	✓	✓		✓			✓*
Relacionamento parcialmente contido			✓				✓*
Integração com SIG			✓				✓
Distinção de pontos de estadia			✓		✓		✓

(*) Apresenta problemas na implementação

Tabela 2.1: Comparação dos modelos analisados.

A análise da tabela permite concluir que, apesar dos trabalhos mencionados, ainda não existe um modelo de dados consistente para análise de comportamento de utilizadores móveis no espaço e no tempo, integrado com informação semântica de forma a aumentar a expressividade do modelo e simplificar a sua compreensão e utilização através de um cubo de dados multidimensional. De facto, o objetivo do modelo proposto é criar as bases para a concretização de aplicações e algoritmos de deteção de comportamentos, atividades de utilizadores móveis, e ainda demonstrar a sua utilidade nas áreas mencionadas.

Na secção seguinte são descritas as diversas características do modelo proposto através do seu processo de modelação dimensional e da modelação do seu processo ETL para enriquecimento semântico dos dados.

Capítulo 3

Data Warehouse Espácio-Temporal

A abordagem comum de armazenamento para a gestão de dados é através de um Sistema de Gestão e Base de Dados (SGBD). Porém, com esta opção a entidade que necessita de gerir os dados fica limitada a processos simples, orientados à transação e a interrogações simplificadas, e os tempos de resposta podem ser extremamente morosos para interrogações mais complexas.

Os sistemas OLAP permitem por parte das empresas, colmatar algumas das limitações dos sistemas OLTP no que diz respeito à gestão de grandes volumes de dados para apoiar os processos de tomada de decisões organizacionais. Portanto, é com naturalidade que quando surge a necessidade de lidar com grandes volumes de dados, e se pretende retirar algum significado dos mesmos, a solução adotada seja um *data warehouse*.

Neste capítulo apresenta-se o modelo para um *data warehouse* espácio-temporal orientado a trajetórias humanas, tal como a modelação semântica das mesmas para uma correta adaptação ao modelo de dados proposto. São ainda apresentadas diversos métodos para o enriquecimento do modelo de dados proposto.

Na Secção 3.1 é feita a modelação dimensional do *data warehouse* espácio-temporal e descrição dos respetivos passos, sendo no final apresentado o modelo proposto; na Secção 3.2 são apresentados métodos de enriquecimento semântico do *data warehouse* aplicadas no processo ETL: (1) caracterização de trajetórias, (2) agrupamento de localizações, (3) extração de pontos de estadia e sua categorização, e (4) algoritmo para descoberta de utilizadores semelhantes.

3.1 Modelação Dimensional

Tal como referido na Secção 2.1.3, a modelação dimensional consiste geralmente em quatro passos, de modo a obter-se um bom modelo dimensional, nomeadamente: (1) listar

prioridades para a construção do *data warehouse*; (2) determinação do nível de detalhe da tabela de factos; (3) modelação das dimensões do negócio; e (4) identificação das medidas numéricas da tabela de factos.

Prioridades para a construção do data warehouse

A prioridade de um DW espácio-temporal orientado a trajetórias é a disponibilização de informação que seja uma mais valia para a utilização nos processos de negócio a que o modelo se propõe ser útil, tais como serviços de planeamento urbano, controlo de tráfego, análise de perfil de utilizadores, *marketing*, entre outros. Portanto, é essencial que o modelo seja centrado na movimentação humana, sendo as coordenadas de GPS o principal foco do modelo, tal como todos os fatores adjacentes, como as características temporais, informação semântica geográfica e perfil do utilizador.

Para além da identificação dos processos de negócio, é fundamental apresentar questões derivadas desses processos e que o DW deverá ter a capacidade de responder. Um dos objetivos do DW proposto para trajetórias humanas, e com base em trabalhos relacionados [4, 6, 17, 32], é a resposta a questões, tais como:

- Qual é o total de utilizadores que se movimentam numa dada região num dado intervalo de tempo?
- Existe alguma diferença substancial na velocidade média dos veículos numa região durante o fim-de-semana?
- Quais os transportes onde os utilizadores passam mais tempo, por períodos do dia, durante a semana e fim de semana?
- Nos bairros da cidade quais os períodos do dia em que existe mais movimentação de veículos e qual a velocidade média de circulação dos mesmos?
- Quais as horas e períodos do dia em que existe mais e menos movimentação (dinâmica da cidade)?
- Para cada período do dia e dia da semana, qual a distribuição de visitas por diferentes utilizadores por ponto de estadia mais visitado?
- Quais as localizações da cidade mais ativas, ou seja, com mais movimentação, em horário laboral aos dias de semana? E em horário pós-laboral?
- Quais as localizações da cidade e pontos de estadia de divertimento noturno com mais movimento?

- Qual é o total de utilizadores que se movimentam numa dada localização e num determinado intervalo de tempo?
- Qual o número de veículos e a sua velocidade média que circulam numa certa localização e num certo dia, por hora?
- Quais os pontos de estadia mais visitados e a duração do tempo de estadia dos utilizadores durante um determinado período de festividades na cidade?

A resposta a este tipo de questões poderá ter utilidade para diversos fins, nomeadamente análise de planeamento urbano, análise de perfil, análise de *marketing*, identificação de padrões espácio-temporais de comportamentos, entre outras. Apesar de as mesmas puderem ser respondidas por sistemas OLTP, a opção recai em OLAP pelos custos computacionais e tempos de resposta serem muito inferiores, assim como a capacidade de análise multidimensional e rápida criação de relatórios analíticos. De facto, estes fatores são críticos para analistas e decisores que utilizem ferramentas de análise de dados.

Determinação do nível de detalhe da tabela de factos

Tendo por base as prioridades enumeradas em cima, podemos então definir a granularidade da tabela de factos do modelo como a representação de um ponto de uma trajetória através das coordenadas capturadas por um certo dispositivo de captura, de uma determinada localização e de um determinado ponto de estadia (pertencente a uma ou mais categorias), relativos a uma trajetória, de um determinado utilizador, num determinado dia e hora, e associadas a um tipo de movimento do utilizador (ver Figura 3.2).

As dimensões necessárias para a constituição da tabela de factos baseiam-se nos requisitos apresentados na Secção 2.1.4, que são identificadas e detalhadas no próximo passo de modelação dimensional.

Modelação das dimensões do negócio

Num processo de modelação dimensional é imperativo efetuar uma construção cuidada das dimensões apuradas. Para atingir as prioridades definidas e corresponder ao nível de detalhe necessário para a tabela de factos, foram consideradas as seguintes dimensões: *Utilizador*, *Trajetória*, *Data*, *Tempo*, *Localização*, *Ponto de Estadia*, *Categoria Ponto de Estadia*, *Dispositivo de Captura*, e *Tipo de Movimento* (ver Figura 3.2). O Anexo B apresenta o detalhe das dimensões, atributos e hierarquias representadas na Figura 3.1.

- Utilizador: dimensão demográfica em que se identifica o utilizador e se caracteriza o perfil do mesmo através de diversos atributos, como por exemplo, sexo, idade, estado civil, entre outros.

- Trajetória: dimensão temática em que se identifica e caracteriza a trajetória correspondente ao ponto da trajetória a que o facto pertence. É nesta dimensão que se faz a caracterização semântica da trajetória, como a distância percorrida, velocidade média e tempo total da trajetória.

- Data: dimensão temporal que identifica a data de um determinado registo, possuindo também informação sobre a data correspondente, como o dia da semana (numérico e textual), se é dia útil, ou ainda se é feriado. Contém a seguinte hierarquia: *Ano* > *EstaçãoAno* > *Mês* > *Quinzena* > *Dia*.

- Tempo: dimensão temporal que identifica a hora e período do dia de um determinado registo. Contém a seguinte hierarquia: *PeriodoDia* > *Hora* > *Minuto* > *Segundo*. Esta dimensão poderia ser representada com a dimensão *Data*, mas esta abordagem iria aumentar em demasia os registos a armazenar no DW. A solução adotada torna os dados desta dimensão sem necessidade de ser alterados. Por exemplo, no caso de um conjunto de dados abranger um período de 365 dias, para a solução de data e tempo na mesma dimensão seriam necessários 31 536 000 registos (365 dias * (24 horas * 60 minutos * 60 segundos)), enquanto que para a solução apresentada são necessários para a dimensão *Data* e *Tempo*, 365 registos e 86,400 registos, respetivamente.

- Localização: dimensão espacial que identifica o setor (região da cidade), ao qual pertencem as coordenadas. Esta dimensão permite adotar a solução de agrupamento de localizações apresentada na Secção 2.1.2.

- Ponto de Estadia: dimensão espacial que obtém uma descrição mais pormenorizada de um ponto de estadia para um determinado registo. Contém a seguinte hierarquia: *Continente* > *Pais* > *Distrito* > *Região* > *Cidade* > *Freguesia* > *Bairro* > *CódigoPostal* > *Rua* > *Numero*.

- Categoria Ponto de Estadia: dimensão temática que identifica as categorias de um ponto de estadia (por exemplo, restaurante, faculdade, entre outras).

- Dispositivo de Captura: dimensão tecnográfica que identifica e caracteriza o dispositivo com que foram obtidas o conjunto de coordenadas do registo.

- Tipo de Movimento: dimensão temática que identifica de que modo o utilizador se estava a deslocar no respetivo registo, e respetiva caracterização desse modo de transporte (por exemplo, bicicleta, táxi, a pé, entre outros).

As dimensões *Localização*, *Ponto de Estadia* e *Utilizador* possuem atributos de registo histórico por se tratarem de dimensões de mudança lenta, isto é, caracterizam-se por serem

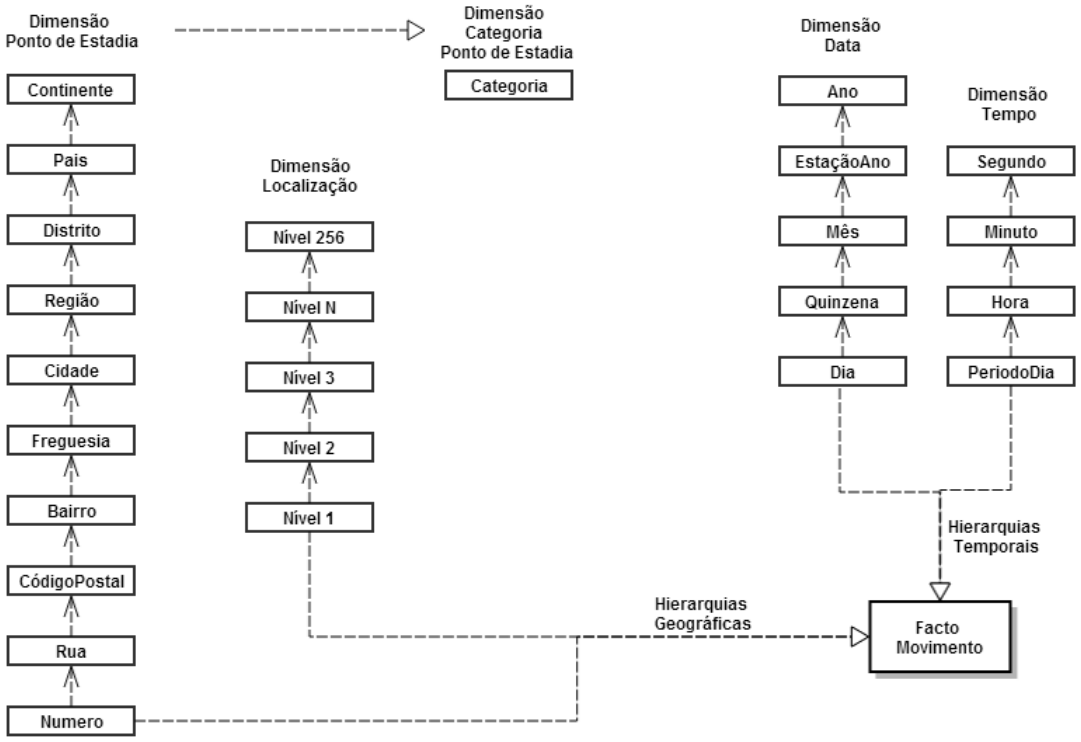


Figura 3.1: Hierarquias presentes no modelo.

atualizadas com menos frequência que as restantes dimensões [26]. Estes atributos são os seguintes: *InicioValidade*, *FimValidade*, *Razao* e *EmVigor*.

Relativamente ao problema do relacionamento parcialmente contido (apresentado em 2.1.2) este não afeta o modelo proposto, pois a representação por localizações e a representação geográfica semântica são feitas, respetivamente, na dimensão *Localização* e dimensão *Ponto de Estadia*. Porém, a resolução deste problema não foi um dos focos deste projeto, sendo um dos possíveis trabalhos futuros.

Outro problema descrito no capítulo anterior é o da contagem distinta. Ao criar identificadores únicos para cada registo de cada dimensão este problema é resolvido, permitindo assim serem efetuadas interrogações que necessitem a distinção de, por exemplo, diferentes utilizadores ou trajetórias. Na Secção 4.3, aquando da implementação do cubo de dados é proposto um aperfeiçoamento através da contagem distinta (*distinct count*) implementada por esta solução.

Por último, devido às diferenças entre os modelos relacionais e modelos analíticos, para a criação da relação entre as dimensões *Ponto de Estadia* e a *Categoria Ponto de Estadia*, foi ainda modelada uma dimensão do tipo ponte, *Ponte Ponto de Estadia*. A modelação deste caso foi baseado no caso de *multivalued dimensions*, referido por Kimball *et al.* [26]. A sua implementação é detalhada aquando da explicação da concretização do cubo de dados (Secção 4.3).

Identificação das medidas numéricas da tabela de factos

A tabela de factos *Movimento* é composta pelas chaves estrangeiras das dimensões e pela chave *OrdemTrajetoria* que corresponde a uma dimensão degenerada [26]. Esta serve apenas para agrupar factos e indicar a ordem das coordenadas pertencentes a uma determinada trajetória, não possuindo assim atributos próprios nem tabela de dimensão associada.

A tabela de factos é ainda composta pela latitude e longitude que caracterizam um ponto da trajetória que faz parte da trajetória, e as medidas numéricas velocidade, aceleração, distância percorrida e tempo decorrido. A descrição detalhada desta tabela encontra-se no Anexo B.

Na Figura 3.2 podemos observar o esquema em estrela para o modelo proposto após a fase de modelação dimensional.

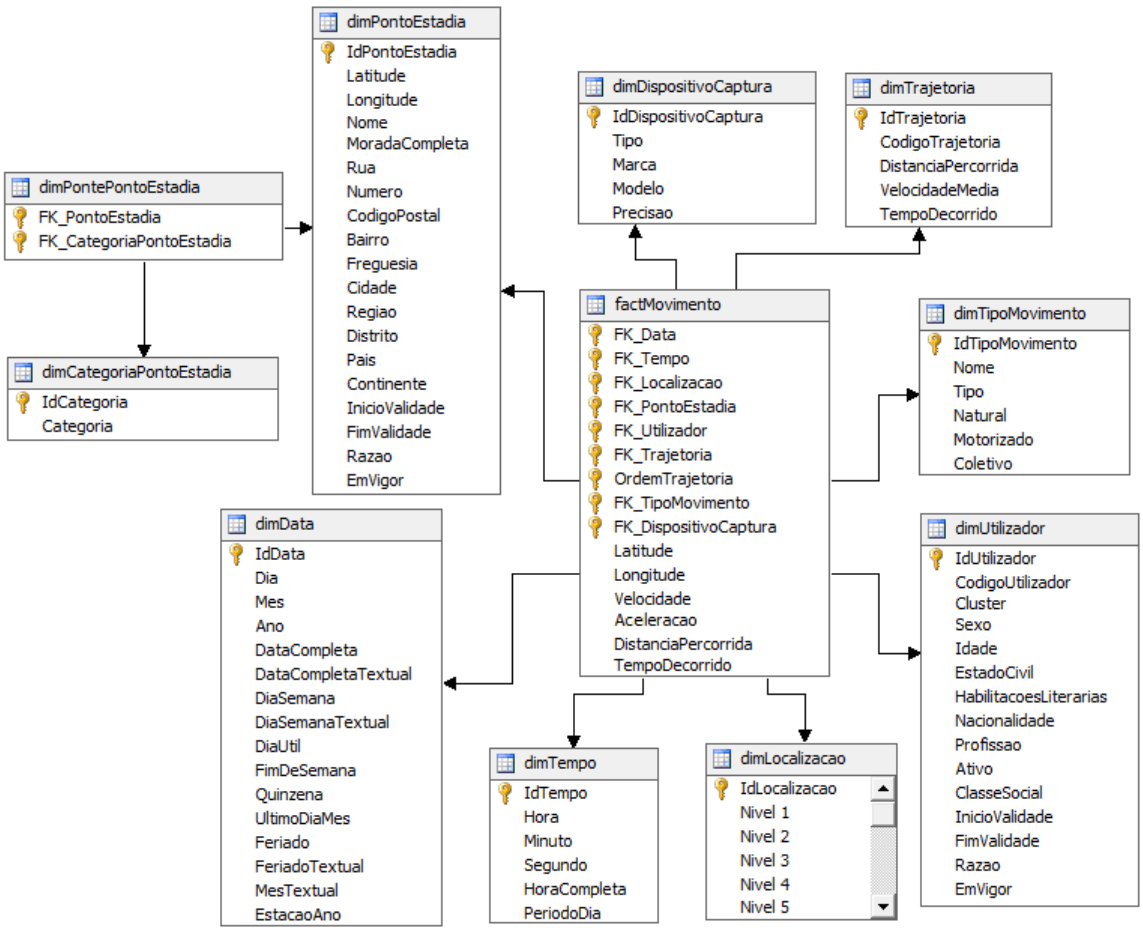


Figura 3.2: Modelo em estrela de representação de trajetórias humanas.

3.2 Modelação do Processo ETL

O processo ETL compreende três etapas [25]: **extração** dos dados do sistema operacional, **transformação** dos dados extraídos tendo em conta as regras de negócio, e o **carregamento** de dados para o DW.

No tratamento dos dados de trajetórias, é essencial que na fase de **transformação** existam métodos definidos para o enriquecimento semântico dos dados. Informações como velocidade média e distância de uma trajetória, e/ou pontos de estadia presentes numa trajetória, são por exemplo importantes para aumentar a utilidade dos dados no DW nas áreas de planeamento urbano e análise de perfil de utilizadores,

O processo de reconstrução de trajetórias é referido na literatura existente como sendo importante no processo ETL [9, 29, 31], porém neste projeto apenas foi dado foco a este processo na fase da concretização do DW, durante o tratamento do conjunto de dados para avaliação.

Serão então apresentadas nas seguintes secções as bases (sem qualquer tipo de concretização e/ou linguagem associadas) para métodos que permitam o enriquecimento semântico do modelo proposto na secção anterior.

3.2.1 Caracterização de Trajetórias

Relativamente à caracterização das trajetórias é possível calcular a velocidade média entre dois pontos seguindo a abordagem de Zheng *et al.* [42]: dados dois pontos consecutivos P_1 e P_2, é calculada a distância D_1, tal como o intervalo temporal t_1, e portanto a velocidade é calculada pela Fórmula 3.1. Para calcular D_1 é utilizada a fórmula de Haversine, já definida na Secção 2.1.2.

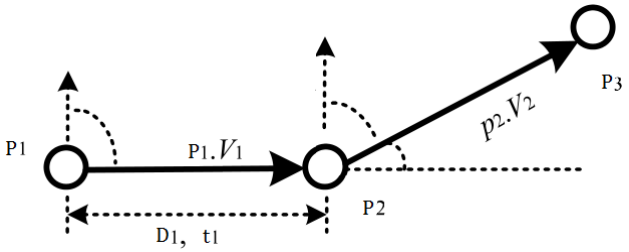


Figura 3.3: Representação de pontos referentes a uma trajetória.

A aceleração é calculada através da Fórmula 3.2, e por fim para calcular t_1 basta apenas efetuar a subtração entre o registo temporal entre P_2 e P_1.

$$P1.V1 = D1/t1 \tag{3.1}$$

$$P1.A1 = P1.V1/t1 \tag{3.2}$$

Para o modelo proposto, esta abordagem permite calcular as medidas *Velocidade*, *Aceleração*, *DistanciaPercorrida* e *TempoDecorrido* referentes à tabela de factos *Movimento*. Em relação aos atributos relativos à dimensão *Trajetória*, para o atributo *DistanciaPercorrida* apenas é necessário calcular a soma das distâncias dos registos de cada trajetória, para o *VelocidadeMedia* calcula-se a média através da divisão do atributo *Velocidade* da tabela *Movimento* e o número de registos da trajetória. Por fim o *TempoDecorrido* é calculado através da soma da medida *TempoDecorrido* dos registos de cada trajetória.

Zheng *et al.* [42] definem ainda um método para o cálculo da direção de cada movimento. Contudo, foi decidido que no processo de modelação esta medida não seria uma mais valia pois a orientação dos movimentos dos utilizadores não se enquadra nos objetivos do modelo.

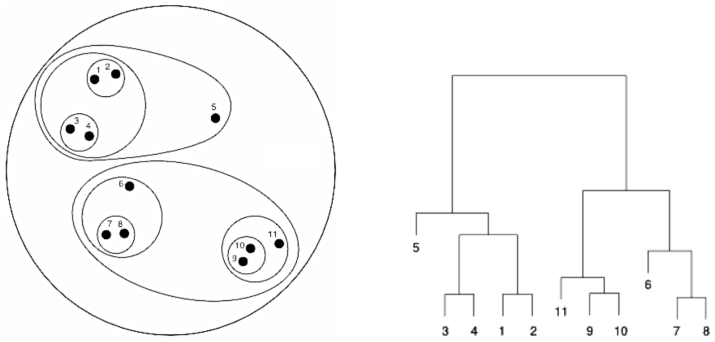
3.2.2 Agrupamento de Localizações

Na Secção 2.1.2 são referidas soluções para a divisão dos registos de trajetórias no espaço geográfico. Neste projeto foi adotado o método de agrupamento, mais concretamente o agrupamento espacial [20]. Esta solução irá permitir criar grupos de registos individuais de trajetórias que estarão divididos por níveis, tal como se pode ver na dimensão *Localização* modelada (ver Anexo B).

Uma abordagem possível para realizar o agrupamento de localizações é a seguinte: a partir de uma amostra do conjunto de registos iniciais (caso o volume de dados seja muito elevado), é efetuado o cálculo das distâncias [21] e aplicado depois um método de agrupamento hierárquico aglomerativo para dividir a amostra inicial em grupos (*clusters*). É utilizado um algoritmo hierárquico pois os registos são assim organizados numa árvore de acordo com a distância entre os mesmos, ficando os registos similares no mesmo ramo.

O funcionamento genérico de um método de agrupamento hierárquico aglomerativo é o seguinte [19]: no primeiro passo, cada elemento do conjunto de dados forma um *cluster*; nas seguintes iterações, são unidos pares de *clusters* que satisfaçam um certo critério de distância mínima, formando cada par um único cluster; o processo termina quando apenas resta um único *cluster* ou quando um *k* número de *clusters* é atingido (ver Figura 3.4).

Após os grupos definidos, o processo permite que seja efetuado um corte da árvore resultante para o número níveis/altura pretendida. No modelo proposto, pode-se observar que foi estimado um máximo de 256 níveis. Este modelo aproxima-se da solução implementada por Braz *et al.* [9] e Orlando *et al.* [31] como pode ser observado na Figura



(a) (b)

Figura 3.4: (a) Resultado final de uma técnica de agrupamento hierárquico aglomerativo (b) Representação em árvore/dendograma dos *clusters* resultantes.

2.13(b), pois a cada nível irá corresponder um conjunto de localizações.

Depois de os grupos finais especificados, é necessário organizar a totalidade dos registos para finalizar o processo. Para esse efeito foi especificado o Algoritmo 1: para cada registo pertencente ao conjunto inicial calcula-se a sua distância euclidiana (ver Secção 2.1.2) relativamente a cada grupo, sendo guardado o grupo para qual o registo tinha uma distância menor. Optou-se pela utilização do algoritmo euclidiano devido à quantidade de cálculos a efetuar (nº de pontos * nª de clusters), apresentando custos computacionais menores que o algoritmo de Haversine.

Algoritmo 1: Algoritmo para distribuição de registos por *clusters*

Input: Um conjunto de pontos $iSet = \{iS\}$ e um conjunto de clusters $clusterSet = \{C\}$
Output: Um conjunto de dados $oSet = \{oS\}$

```
1  foreach iSet iS do
2      mindist = 1e34;
3      mincluster = −1;
4      foreach clusterSet c do
5          dx = iS.x − c.x;
6          dy = iS.y–c.y;
7          dist = dx ∗ dx + dy ∗ dy;
8          if dist < mindist then
9              mindist = dist;
10             mincluster = c.id;
11     oS.x = iS.x;
12     oS.y = iS.y;
13     oS.idGrupo = c.id;
14     oSet.insert(oS);
15 return oSet;
```

3.2.3 Extração de Pontos de Estadia

Um ponto de estadia, tal como já definido na Secção 2.1.1, corresponde a uma região geográfica onde o utilizador permaneceu durante um certo período de tempo. O processo de extração de pontos de estadia de trajetórias é muito importante, pois pode significar que um utilizador móvel teve algum comportamento significativo naquela localização, podendo revelar assim os seus interesses (por exemplo, fazer compras, ver um filme ou visitar um museu). Ter conhecimento destes pontos de estadia permite uma melhor compreensão dos interesses de utilizador móvel, permitindo também relacionar diferentes utilizadores [45].

Para a extração dos pontos de estadia de trajetórias, foi elaborado o Algoritmo 2 adaptado da metodologia seguida em algumas propostas presentes na literatura [27, 44, 45]. O algoritmo estabelece dois limites, *distLim* e *tempLim*, que definem respetivamente, o limite máximo de distância a que um certo ponto P_i pode estar do ponto inicial da região, e o limite máximo de tempo que um utilizador fica numa determinada área.

Algoritmo 2: Algoritmo para extração pontos de estadia

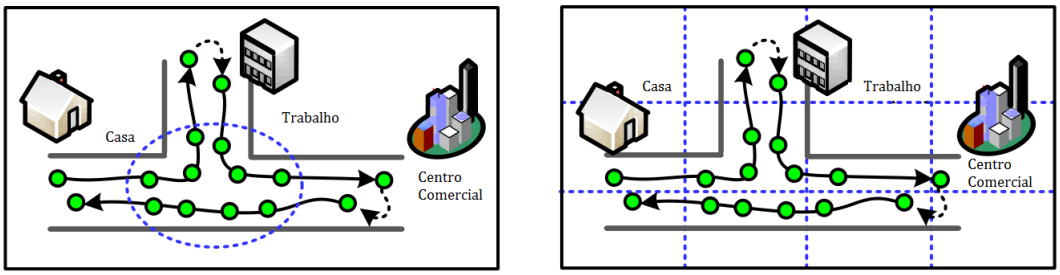
Input: Uma Trajetória T, um limite de distância $distLim$, e um limite de tempo $tempLim$
Output: Um conjunto de pontos de estadia $PE = \{P\}$

```
1  i = 0, numPontos = |T|;
2  while i < numPontos = 0 do
3      j = j + 1, Token = 0;
4      while j < numPontos = 0 do
5          dist = Distancia(p_i, p_j);
6          if dist > distLim then
7              tempDecorrido = p_j.T − p_i.T;
8              if tempDecorrido > tempLim then
9                  P.coordMedia = calculaCoordMedia({p_k | i ≤ k ≤ j});
10                 PE.inserir(P);
11                 i = j, Token = 1;
12             break;
13         j = j + 1;
14     if Token! = 1 then i = i + 1;
15 return PE;
```

Para adicionar informações semânticas sobre cada ponto de estadia, nomeadamente as encontradas na dimensão *Ponto de Estadia* do modelo na Figura 3.2, é utilizado o serviço Google Geocoding API[1] (mais precisamente o serviço de Reverse Geocoding[2]).

[1] https://developers.google.com/maps/documentation/geocoding/
[2] https://developers.google.com/maps/documentation/geocoding/#ReverseGeocoding

Apesar de existirem outras alternativas para a deteção de pontos de estadia [27, 45], nomeadamente agrupamento espacial (por exemplo, DBSCAN), optou-se pelo método anterior na medida em que, com técnicas de agrupamento situações como a entrada e saída de um utilizador num edifício não iriam ser consideradas pontos de estadia uma vez que não existiriam pontos suficientes para formar agrupamentos nestes algoritmos (Figura 3.5(a)). Outra possibilidade seria através de métodos de partição por grelhas regulares (Figura 3.5(b)). Porém também existiriam falhas na deteção de pontos de estadia, pois os casos em que existissem poucos pontos numa célula (por exemplo, entrada e permanência dentro de um edifício), não seriam considerados pontos de estadia.



(a) (b)

Figura 3.5: Outras técnicas de deteção de pontos de estadia. (a) Deteção por agrupamento (b) Deteção por grelha regular.

Por fim, a abordagem utilizada para a categorização de pontos de estadia é feita através do Google Places[3], que permite obter informação mais pormenorizada sobre um par de coordenadas, sendo que para o modelo o interesse recaí sob as categorias e o respetivo nome do local, se disponível.

3.2.4 Descoberta de Utilizadores Semelhantes

Um dos principais objetivos deste projeto é a análise de perfil de utilizadores móveis. Através de grandes volumes de dados sobre o posicionamento de utilizadores, é potenciado o conhecimento sobre as atividades, preferências, padrões de comportamento e de mobilidade desses utilizadores no espaço e ao longo do tempo. Porém, na generalidade dos conjuntos de dados de trajetórias apenas é disponibilizado o conteúdo espacial e temporal das mesmas, sendo dada primazia à privacidade do perfil dos utilizadores, não existindo por vezes sequer um identificador numérico dos mesmos.

Não existindo esta informação sobre os utilizadores foi desenvolvido um método para a descoberta de utilizadores semelhantes através dos pontos de estadia e localizações frequentadas, sendo estas duas características previamente calculadas através dos métodos

[3]https://developers.google.com/places/documentation/

anteriormente apresentadas. Neste método serão considerados utilizadores semelhantes aqueles que visitem os mesmos pontos de estadia e que frequentem as mesmas localizações, criando no fim do processo grupos de utilizadores em que o grau de semelhança é maior.

1ª Fase: Nesta primeira fase é calculada para cada utilizador a frequência com que este visita cada ponto de estadia e localização. Isto é feito através do cálculo da probabilidade de cada utilizador visitar cada ponto de estadia e cada localização. Por exemplo, sabendo que um certo utilizador visitou quatro vezes o ponto de estadia X, uma vez Y, e uma vez Z, a distribuição será aproximadamente P(X) = 0.66, P(Y) = 0.16 e P(Z) = 0.16. No final, cada utilizador tem uma lista com as probabilidades de frequentar cada ponto de estadia, e uma segunda lista com as probabilidades de frequentar cada localização.

Assumindo que um ponto de estadia e/ou localização é mais relevante do que outro(s) quando aparece em mais trajetórias de um determinado utilizador (por exemplo, um utilizador pode visitar diversas vezes um ponto de estadia, mas apenas numa única trajetória), então a relevância de um ponto de estadia ou localização *c* de cada utilizador pode ser calculada, através da divisão do número de trajetórias em que aparece *c* pelo número total de trajetórias do utilizador (Fórmula 3.3).

$$Relevancia(c) = \frac{nPresencaTrajetorias}{totalTrajetoriasUtilizador} \tag{3.3}$$

2ª Fase: Na segunda fase, para cada utilizador é analisada a lista de pontos de estadia e localizações e comparada com todos os outros utilizadores. Quando um utilizador tem um ponto de estadia ou localização igual a outro utilizador, é guardado o identificador desse utilizador e ponto de estadia ou localização idêntica.

3ª Fase: Na terceira fase são calculados os valores de semelhança entre cada utilizador. Utilizadores que visitaram um ponto de estadia que tenha poucas visitas, têm uma maior probabilidade de estarem relacionados, do que se partilhassem um ponto de estadia que é frequentado por muitos utilizadores. Para modelar esta assunção é utilizada a Fórmula 3.4 (*visited popularity*) [27, 45], em que *N* representa o número total de utilizadores e *n* o número de utilizadores distintos que visitam o ponto de estadia ou localização *c*, sendo uma função exponencial, o resultado será que um ponto de estadia menos visitado tem um maior peso no valor de semelhança entre utilizadores.

$$vp(c) = \log \frac{N}{n} \tag{3.4}$$

Para o cálculo dos valores de semelhança entre utilizadores é utilizada a Fórmula 3.5: para cada ponto de estadia c, é calculada a soma dos valores dos pontos de estadia (e posteriormente para as localizações) que o utilizador 1 e 2 (*Util1* e *Util2*) têm em comum, sendo este valor dividido pelo total de pontos de estadia/localizações em comum.

$$sem(Util1, Util2) = \frac{\sum vp(c)_i}{totalNc} \tag{3.5}$$

Para concluir este processo, é feita a união entre os valores de semelhança dos pontos de estadia e localizações em comum entre cada utilizador. Estes valores foram calculados em separado, pois podemos fazer a assunção que dois utilizadores frequentarem o mesmo ponto de estadia é menos provável (ou seja, mais relevante) do que os dois frequentarem a mesma localização do espaço geográfico. No processo de união dos valores, é atribuído o peso de 0.75 aos valores dos pontos de estadia e 0.25 às localizações para representar o grau de importância de cada um dos atributos no processo. Na Figura 3.6 estão representados os diversos passos envolvidos no cálculo de utilizadores semelhantes.

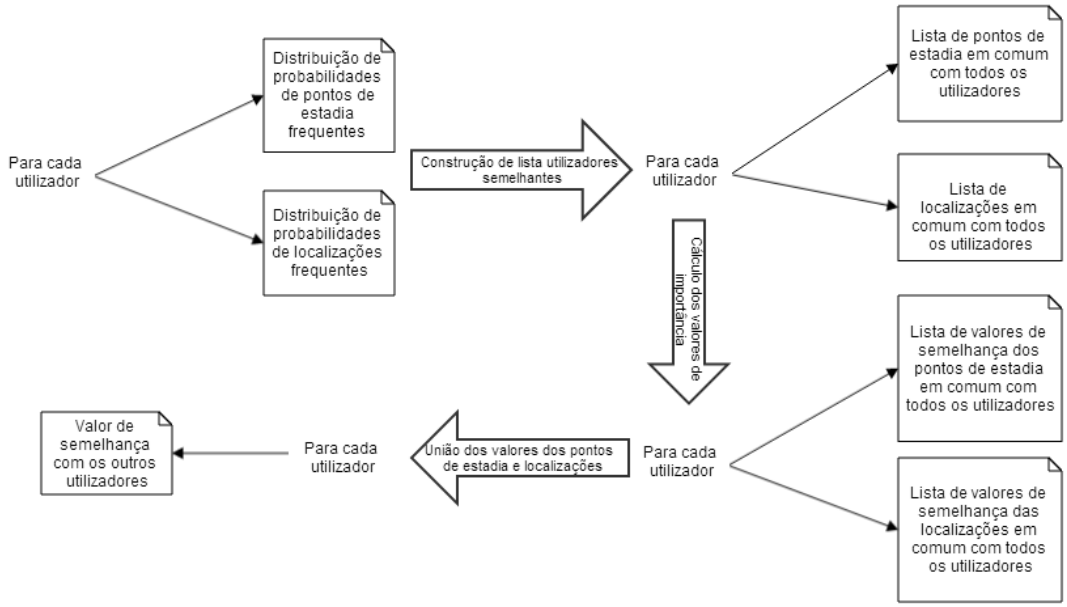


Figura 3.6: Método de cálculo de utilizadores semelhantes.

Como referido anteriormente, através dos resultados obtidos por este processo é possível criar grupos de utilizadores semelhantes. Através da técnica de agrupamento aglomerativo hierárquico, é possível criar grupos em que o perfil de locais visitados sejam idênticos. Desta forma é possível criar informação útil que se engloba na área de análise de perfil de utilizadores e também contribuir para a área de análise de *marketing* (por exemplo, quais os grupos de utilizadores que frequentam os pontos de estadia x e y?). A identificação do grupo de um utilizador está representado no atributo *cluster* na dimensão *Utilizador* do modelo do DW (ver Figura 3.2).

Capítulo 4

Validação do Modelo Proposto

Neste capítulo é apresentada a concretização do modelo proposto, tal como a sua validação. O modelo é concretizado através da sua aplicação a um conjunto de dados de trajetórias composta por 182 utilizadores num espaço temporal de 5 anos. A aplicação dos dados foi efetuada mediante um processo ETL, no qual foram aplicados os métodos desenvolvidos no capítulo anterior. Para a validação da concretização do DW, são realizadas diversas interrogações relacionadas com os processos de negócio propostos como exemplo. É ainda efetuada a demonstração da interligação do DW com uma ferramenta de visualização desenvolvida para este propósito. É ainda realizada a validação do método de descoberta de utilizadores semelhantes.

Este capítulo está então organizado da seguinte forma: na Secção 4.1 é apresentado o conjunto de dados relativos ao projeto Geolife; na Secção 4.2 é apresentada a concretização do processo ETL relativa ao modelo proposto tal como uma análise da dimensão do DW; na Secção 4.3 é descrita a implementação do cubo de dados e por fim na Secção 4.4 é realizada a experimentação do modelo.

4.1 Conjunto de Dados Geolife

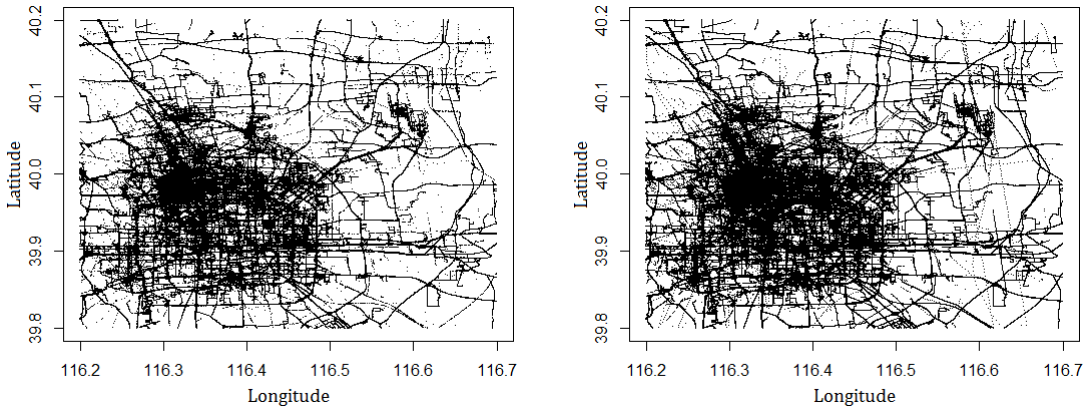
O conjunto de dados utilizado neste processo foi disponibilizado pela *Microsoft Research Asia* relativos ao projeto Geolife [41]. Os dados correspondem à versão 1.3 (Tabela 4.1) do conjunto de dados, sendo recolhidos por um total de 182 utilizadores durante um período de 5 anos (de Abril de 2007 a Agosto de 2012). Os dados foram recolhidos através de sensores e telemóveis com GPS estando sujeito a erros, como qualquer dispositivo com GPS. Os pontos das trajetórias estão registados com uma marca temporal de 1 a 5 segundos ou espacial de 5 a 10 metros entre pontos.

Apesar dos dados abrangerem cerca de 30 cidades na China e existirem registos em outros países/continentes, para este projeto foram apenas utilizados os dados em Beijing,

	Version 1.2	Version 1.3	Change
Time span of the collection	04/2007 – 10/2011	04/2007 – 8/2012	+10 months
Number of users	178	182	+4
Number of trajectories	17,621	18,670	+1,049
Number of points	23,667,828	24,876,978	+1,209,150
Total distance	1,251,654km	1,292,951km	+41,297 km
Total duration	48,203hour	50,176hour	+1,973 hour
Effective days	10,413	11,129	+716

Tabela 4.1: Detalhes do conjunto de dados Geolife [41].

dado que das 18 670 trajetórias do conjunto de dados 17 107 estão concentradas nesta cidade. Nas Figuras 4.1(a) e 4.1(b) podem-se observar graficamente os dados.



(a) (b)

Figura 4.1: (a) Representação de 25% dos dados (b) Representação de 100% dos dados.

Os dados recolhidos do GPS estão disponíveis em milhares de ficheiros de formato PLT[1] (ver formato na Figura 4.3), sendo que cada ficheiro representa uma trajetória de um dado utilizador. Os principais campos de cada entrada dos ficheiros são a latitude e longitude do ponto de GPS (em graus decimais), altitude e data e tempo no formato AAAA/MM/DD HH:MM:SS. Existem ainda utilizadores que têm um ficheiro 'labels.txt' com entradas relativas à caracterização de diversos tipos de movimentação (terrestres, aéreos ou marítimos) dos seus registos, cujo detalhe se pode observar na Tabela 4.2.

4.2 Processo ETL

Tal como já referido em secções anteriores, o processo ETL compreende três etapas [25]: **extração** dos dados do sistema operacional, **transformação** dos dados extraídos tendo em conta as regras de negócio, e o **carregamento** de dados para o DW. Uma correta concretização deste processo é fulcral para o DW, porém dado que o ciclo de vida de um

[1] http://www.rus-roads.ru/gps/help_ozi/fileformats.html

Transportation mode	Distance (km)	Duration (hour)
Walk	10,123	5,460
Bike	6,495	2,410
Bus	20,281	1,507
Car & taxi	32,866	2,384
Train	36,253	745
Airplane	24,789	40
Other	9,493	404
Total	14,0304	12,953

Tabela 4.2: Distância e duração total dos modos de transporte [41].

DW é um processo evolutivo, é durante as fases de implementação e exploração do cubo de dados que vão ser descobertos erros, sendo necessário realizar o processo ETL novamente para a atualização do DW. Na Figura 4.2 podemos observar os principais passos que foram seguidos neste processo e que serão descritos nas secções seguintes.

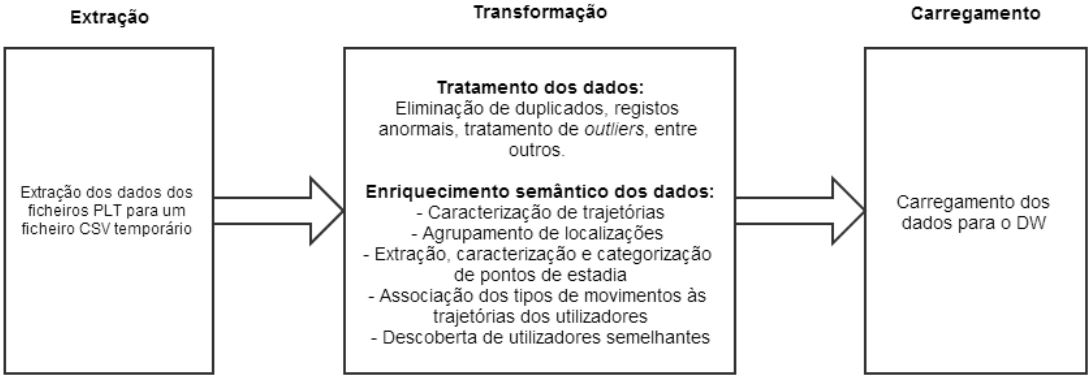


Figura 4.2: Planeamento do processo ETL.

4.2.1 Extração

A fase de extração tem como objetivo ler os dados do sistema OLTP e extrai-los para a *staging area*. A extração foi efetuada a partir dos ficheiros PLT (Figura 4.3), sendo realizada uma análise a partir do domínio e das regras de integridade das colunas, tal como a qualidade e a quantidade de dados disponíveis para a respetiva extração.

Seguindo um processo adaptado ao apresentado no sítio eLifeLog [1], os dados foram importados dos ficheiros de origem para uma base de dados temporária implementada em MySQL[2] através de um *script* PHP[3]. Para terminar o processo de extração, os dados foram exportados da base de dados para um único ficheiro CSV[4], com o seguinte formato

[2]http://www.mysql.com/
[3]http://www.php.net/
[4]http://www.digitalpreservation.gov/formats/fdd/fdd000323.shtml

Line 1…6 are useless in this dataset, and can be ignored. Points are described in following lines, one for each line.

Field 1: Latitude in decimal degrees.

Field 2: Longitude in decimal degrees.

Field 3: All set to 0 for this dataset.

Field 4: Altitude in feet (-777 if not valid).

Field 5: Date - number of days (with fractional part) that have passed since 12/30/1899.

Field 6: Date as a string.

Field 7: Time as a string.

Note that field 5 and field 6&7 represent the same date/time in this dataset. You may use either of them.

Example:

39.906631,116.385564,0,492,40097.5864583333,2009-10-11,14:04:30

39.906554,116.385625,0,492,40097.5865162037,2009-10-11,14:04:35

Figura 4.3: Formato dos ficheiros PLT [41].

para posterior limpeza e transformação: *userid, trajetória, ordem_trajetoria, latitude, longitude, altitude, gps_UTC_timestamp*. A Tabela 4.3 representa de forma parcial os dados desta tabela.

"1"	"0"	"39.984702"	"116.318417"	"0"	"492"	"2008-10-23 02:53:04"	"1224726784"
"1"	"0"	"39.984683"	"116.31845"	"0"	"492"	"2008-10-23 02:53:10"	"1224726790"
"1"	"0"	"39.984686"	"116.318417"	"0"	"492"	"2008-10-23 02:53:15"	"1224726795"
"1"	"0"	"39.984688"	"116.318385"	"0"	"492"	"2008-10-23 02:53:20"	"1224726800"
"1"	"0"	"39.984655"	"116.318263"	"0"	"492"	"2008-10-23 02:53:25"	"1224726805"
"1"	"0"	"39.984611"	"116.318026"	"0"	"493"	"2008-10-23 02:53:30"	"1224726810"
"1"	"0"	"39.984608"	"116.317761"	"0"	"493"	"2008-10-23 02:53:35"	"1224726815"
"1"	"0"	"39.984563"	"116.317517"	"0"	"496"	"2008-10-23 02:53:40"	"1224726820"
"1"	"0"	"39.984539"	"116.317294"	"0"	"500"	"2008-10-23 02:53:45"	"1224726825"
"1"	"0"	"39.984606"	"116.317065"	"0"	"505"	"2008-10-23 02:53:50"	"1224726830"
"1"	"0"	"39.984568"	"116.316911"	"0"	"510"	"2008-10-23 02:53:55"	"1224726835"
"1"	"0"	"39.984586"	"116.316716"	"0"	"515"	"2008-10-23 02:54:00"	"1224726840"

Tabela 4.3: Fragmento exemplificativo da tabela temporária.

4.2.2 Transformação

Após a etapa de extração segue-se a etapa de transformação (Figura 4.2). Esta etapa envolveu dois processos: o tratamento de dados e o enriquecimento semântico dos dados. No primeiro foi efetuado os processos básicos desta etapa como a verificação de dados inválidos e tratamento de *outliers*. No enriquecimento foram aplicados os métodos de extração do tipo de movimentos de utilizadores e o enriquecimento de diversas dimensões do modelo.

Tratamento dos dados:

Esta fase foi composta por diversas tarefas (tal como, eliminação de duplicados, tratamento de *outliers*, correção de registos temporais, entre outros) e foi feita em diversos ciclos, sendo algumas das tarefas realizadas após a concretização do DW. Isto deveu-se

ao facto de surgirem novos requisitos ao longo do projeto, o que levou à criação de novas regras de negócio e à necessidade da criação e respetivo tratamento de nova informação. As principais tarefas de tratamento de dados foram realizadas na linguagem Python[5] e são as seguintes:

- Eliminação de dados duplicados: para os registos duplicados que eram caracterizados por pertencerem ao mesmo utilizador, mesma trajetória, e possuindo o mesmo par de coordenadas e registo temporal, foi efetuada a eliminação de dados.

- Eliminação de *outliers* (definidos na Secção 2.1.1): esta tarefa foi feita através da análise das trajetórias, comparando as distâncias (método de Haversine) entre os seus registos. O primeiro passo envolveu o cálculo da distância média percorrida entre os pontos de cada trajetória; no segundo passo foi realizada a análise da distância entre pontos de cada trajetória, sendo que se a distância entre P_i e P_{i+1} fosse superior à média da trajetória, P_{i+1} seria um possível *outlier*. Para melhorar o processo foi efetuada a comparação com a distância de P_i a P_{i+1} com P_{i+1} a P_{i+2} para evitar situações em que o tipo de movimento foi alterado, e nestes casos a distância entre pontos ser maior (por exemplo, o utilizador estiver a andar a pé e entrar num comboio, ou seja, a velocidade do utilizador aumenta o que provoca registos com mais distância entre eles).

- Eliminação de marcas erradas: foram encontrados também diversos registos com marcas temporais erradas (por exemplo, referentes ao ano 2000), sendo estes registos eliminados.

- Correção de fuso horário: por último, foi alterada o registo temporal de todos os dados pois o conjunto de dados estava uniformizado no fuso horário GMT (*Greenwich Mean Time*) mas tendo em conta que apenas estamos a lidar com dados na área geográfica de Beijing, os dados foram alterados para o fuso horário em vigor, ou seja, de GMT para GMT+8.

Enriquecimento semântico dos dados:

Se o passo anterior foi essencial para assegurar a validade dos dados do DW, este passo é essencial para garantir a sua qualidade.

A primeira tarefa envolveu o enriquecimento de duas dimensões temporais (através do Microsoft Excel[6]): na dimensão *Tempo* foram caracterizados os períodos do dia, e na dimensão *Data* foram preenchidos os atributos de cada registo, tais como atributos

[5]http://www.python.org/
[6]http://office.microsoft.com/en-us/excel/

textuais, por exemplo se é um dia útil, se é feriado, sempre tendo em conta o calendário chinês em vigor no ano correspondente.

A segunda tarefa envolveu a extração dos tipos de movimento dos utilizadores através dos ficheiros 'labels.txt' já referidos. Os registos dos tipos de movimentos efetuados pelos utilizadores estavam associados apenas por marcas temporais (formato: *Start Time, End Time, Transportation Mode*), não tendo qualquer trajetória associada. O modelo proposto neste projeto é centrado no movimento do utilizador e não propriamente na trajetória, o que facilitou este processo, envolvendo apenas a comparação (através de um *script* em Python) dos registos temporais dos movimentos dos utilizadores e atribuindo o respetivo identificador do tipo de movimento. Aos registos para os quais não foi possível associar um tipo de movimento foi atribuída a designação de *desconhecido*.

Por fim, foram aplicados os métodos desenvolvidos na Secção 3.2. Através da abordagem de caracterização de trajetórias, foi calculada a distância percorrida (em metros), o tempo decorrido (em segundos), e a média da velocidade para cada trajetória, sendo também calculados estes atributos para cada movimento (medidas numéricas definidas na tabela de factos).

Dado que Beijing é uma região para a qual não foi possível obter informação sobre a divisão administrativa em subregiões ou bairros, foi aplicado o método de agrupamento de localizações ao conjunto de dados. Aplicando o processo especificado na Secção 3.2.2 em R[7] e através de uma amostra de 10% dos dados, foi obtida a árvore com os respetivos registos, utilizando o método hierárquico aglomerativo *ward* [19]. Dado que não foi especificado o número de grupos final, a árvore foi gerada com o número de grupos final correspondente ao número de registos da amostra inicial (aproximadamente 2 milhões). Para efetuar o corte da árvore para o número de grupos desejado foi utilizado o comando *cuttree* [21] obtendo-se assim uma árvore com apenas 256 grupos. Este número corresponde ao número de níveis da dimensão *Localização*. Na Figura 4.4 podemos observar uma pequena amostra com apenas 21 grupos.

Para efetuar a extração dos pontos de estadia, foi aplicado o Algoritmo 2 sendo que os limites usados foram: para *tempLim* 30 minutos e para *distLim* 200 metros, sendo obtidos cerca de 21 000 pontos de estadia. Neste processo foram detetados milhares de pontos de estadia que pertenciam ao mesmo local. De modo a resolver este problema, foi realizada uma análise dos dados através do atributo *MoradaCompleta* para eliminar os pontos de estadia repetidos e reduzir o seu número para análise. Através desta operação obteve-se cerca de 4 900 pontos de estadia.

Para efetuar a caracterização dos pontos de estadia foi utilizada a biblioteca Pyge-

[7] http://www.r-project.org/

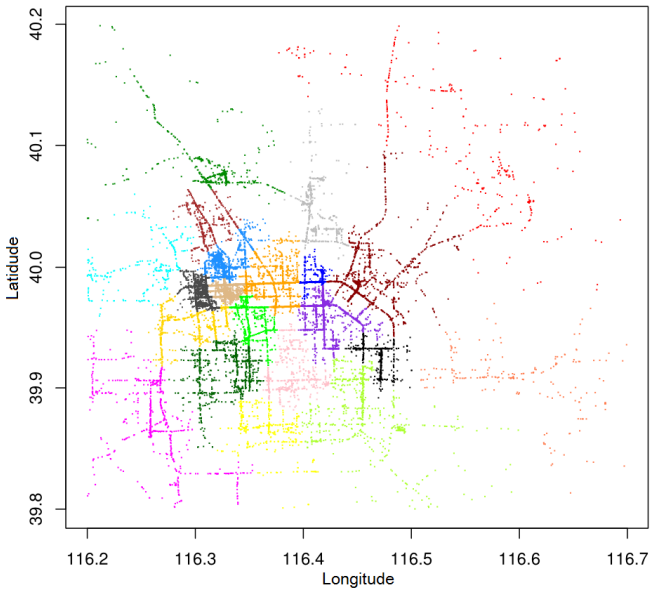


Figura 4.4: Exemplo da amostra de dados com 21 grupos (cada cor representa um grupo).

ocoder[8] em Python que utiliza o já referido Google Geocoding para obter as diversas informações modeladas na dimensão *Ponto de Estadia*. Por fim, para efetuar a categorização foi usada a biblioteca Google APIs Client Library[9] em Java[10] que permite utilizar o *web service* Google Places.

Para terminar esta etapa, foi aplicado o algoritmo de descoberta de utilizadores semelhantes. Para cada utilizador foi obtido um ficheiro que continha os valores de semelhança com todos os outros utilizadores. Para efetuar a criação de grupos de utilizadores semelhantes foi necessário criar uma matriz com todos os valores de semelhança, sendo que estes foram transformados para obter os valores de distância (Fórmula 4.1) necessários ao algoritmo de agrupamento.

$$Distancia(v) = -\log(v + \epsilon(1 - v)) \tag{4.1}$$

Na Tabela 4.4 mostra-se a matriz parcial resultante, podendo-se observar a distancia 0.0 na comparação com o próprio utilizador. Após a criação da matriz, foi realizado em R o carregamento da matriz e definido o espaço multi-dimensional através do comando *cmdscale*[11]. Após este passo foi feito o agrupamento de utilizadores através da técnica *ward*, resultando num dendograma (ver Figura D.1, no Anexo D) que depois de analisada, resultou em 9 grupos de utilizadores.

[8] https://bitbucket.org/xster/pygeocoder/wiki/Home
[9] https://code.google.com/p/google-api-java-client/
[10] http://www.java.com/en/download/faq/whatis_java.xml
[11] http://stat.ethz.ch/R-manual/R-patched/library/stats/html/cmdscale.html

	U001	U002	U003	U004	U005	U006
U001	0.000	0.758	0.582	0.356	0.526	0.655
U002	0.758	0.000	0.658	0.770	0.711	0.857
U003	0.582	0.658	0.000	0.580	0.566	0.617
U004	0.356	0.770	0.580	0.000	0.475	0.597
U005	0.526	0.711	0.566	0.475	0.000	0.672
U006	0.655	0.857	0.617	0.597	0.672	0.000

Tabela 4.4: Matriz parcial com valores de utilizadores semelhantes.

Como exemplo do resultado desta fase de transformação, podemos observar a Tabela 4.5 que mostra (parcialmente) o aspeto da tabela de factos *Movimento* no respetivo CSV.

```
523;39291;77;0;1;1;22;0;0;39,984532000000000;116,314808000000000;1,82238560890397850000;9,11192804452000000000;5
523;39296;77;0;1;1;23;0;0;39,984504000000000;116,314625000000000;3,17986876170047730000;15,89934380850000000000;5
523;39301;77;0;1;1;24;0;0;39,984485000000000;116,314426000000000;3,41716858131950470000;17,08584290660000000000;5
523;39306;77;0;1;1;25;0;0;39,984427000000000;116,314240000000000;3,42184271390135200000;17,10921356950000000000;5
523;39311;77;0;1;1;26;0;0;39,984485000000000;116,314042000000000;3,61206071438615960000;18,06030357190000000000;5
523;39316;77;0;1;1;27;0;0;39,984480000000000;116,313818000000000;3,81856185251511440000;19,09280926260000000000;5
523;39321;77;0;1;1;28;0;0;39,984501000000000;116,313659000000000;2,74930339720473870000;13,74651698600000000000;5
523;39326;77;0;1;1;29;0;0;39,984618000000000;116,314323000000000;11,60982387287731800000;58,04911936440000000000;5
523;39331;77;0;1;1;30;0;0;39,984649000000000;116,314107000000000;3,74462429220870160000;18,72312146100000000000;5
523;39336;77;0;1;1;31;0;0;39,984621000000000;116,313941000000000;2,89634944748369530000;14,48174723740000000000;5
523;39341;77;0;1;1;32;0;0;39,984655000000000;116,313724000000000;3,77417219902497800000;18,87086099510000000000;5
523;39346;77;0;1;1;33;0;0;39,984681000000000;116,313521000000000;3,50708820572689370000;17,53544102860000000000;5
523;39351;77;0;1;1;34;0;0;39,984708000000000;116,313311000000000;3,62840091974476530000;18,14200459870000000000;5
523;39356;77;0;1;1;35;0;0;39,984708000000000;116,313099000000000;3,61245154646907200000;18,06225773230000000000;5
523;39361;133;0;1;1;36;0;0;39,984696000000000;116,312921000000000;3,04481396165608050000;15,22406980830000000000;5
523;39366;133;0;1;1;37;0;0;39,984677000000000;116,312746000000000;3,01176529263466450000;15,05882646320000000000;5
523;39371;133;0;1;1;38;0;0;39,984682000000000;116,312525000000000;3,76745320474602740000;18,83726602370000000000;5
523;39376;133;0;1;1;39;0;0;39,984649000000000;116,312332000000000;3,36958628530497100000;16,84793142650000000000;5
523;39381;133;0;1;1;40;0;0;39,984641000000000;116,312123000000000;3,56577639231964200000;17,82888196160000000000;5
```

Tabela 4.5: Tabela de factos *Movimento* parcial.

As tabelas referentes às dimensões que resultaram do processo de transformação, encontram-se no Anexo C.

4.2.3 Carregamento

Nesta última fase do processo ETL foi efetuado o carregamento de dados para o DW. Este processo foi efetuado utilizando a ferramenta *Integration Services* [28] do SQL Server, que permite automatizar os processos e efetuar o tratamento de erros que possam surgir durante o carregamento de dados, tais como incompatibilidades de formatos presentes nos ficheiros de entrada e as colunas das tabelas no SQL Server. Para tal, foi concebido um processo iterativo para efetuar a ligação entre os ficheiros CSV resultantes da fase de transformação e as tabelas criadas anteriormente no SQL Server.

Para efetuar o carregamento das dimensões *Ponto de Estadia*, *Categoria Ponto de Estadia* e respetiva dimensão de ponte, foi criado um outro processo, porque o facto da dimensão de ponte (*Ponte Ponto de Estadia*) ser constituída por chaves estrangeiras, as dimensões referenciadas tiveram de ser carregadas em primeiro lugar. Na Figura 4.5 pode-se observar com mais pormenor o processo ETL aplicado.

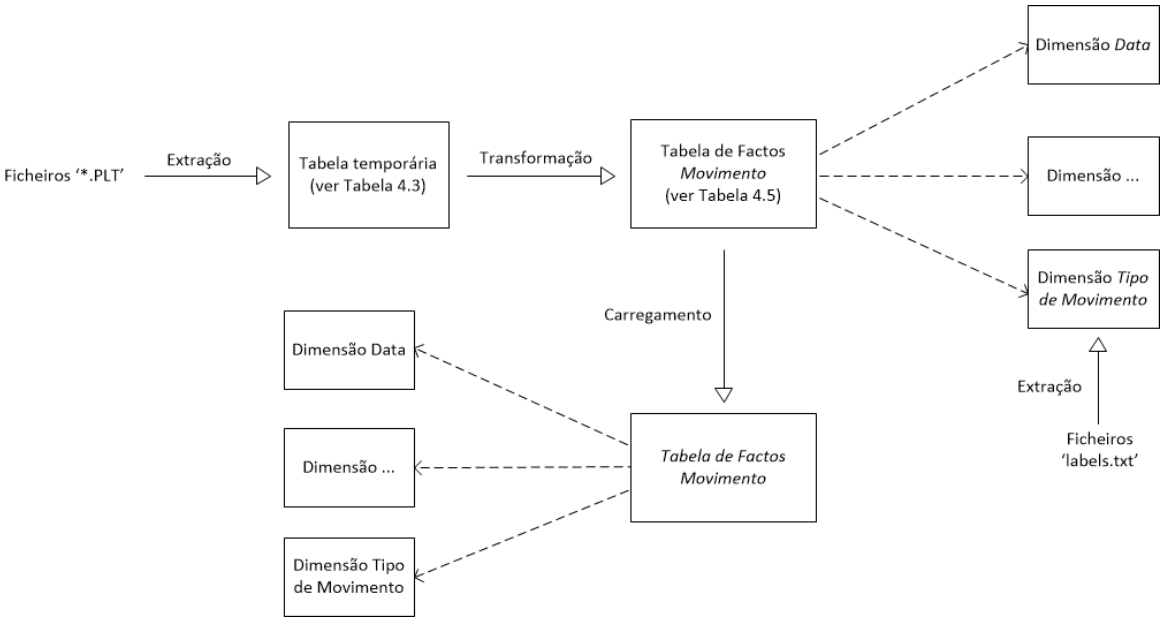


Figura 4.5: Processo ETL aplicado.

4.2.4 Análise da Dimensão do Data Warehouse

Nesta secção apresenta-se uma pequena análise do impacto físico do conjunto de dados aplicado ao modelo proposto. Esta informação é gerada através dos relatórios que o SQL Server permite gerar em relação ao DW alojado no servidor. A informação apresentada corresponde, como já referido anteriormente, aos dados localizados apenas em Beijing, abrangendo um total de 17 107 trajetórias, 179 utilizadores com registos espalhados por por um período temporal de 5 anos.

Na Tabela 4.6 podemos observar o espaço ocupado por cada dimensão e respetiva tabela de factos. Como é habitual num DW, a tabela de factos possui uma quantidade de registos incomparavelmente superior às restantes dimensões (17 704 084 registos) e consequentemente ocupam um maior espaço em disco. As restantes dimensões têm um tamanho aceitável, sendo também normal que as dimensões temporais ocupem bastante espaço em disco.

Na Tabela 4.7 podemos observar o tamanho real do DW, cerca de 18 255.69 MB (*Megabytes*) divididos pelos ficheiros de dados das tabelas criadas (2 637.00 MB) e o ficheiro de registo de transações do DW (15 618.69 MB). Este último é o responsável pelo aumento de espaço em disco, embora seja também aquele que tem maior importância pois contém o registo de todas as transações efetuadas sob o DW.

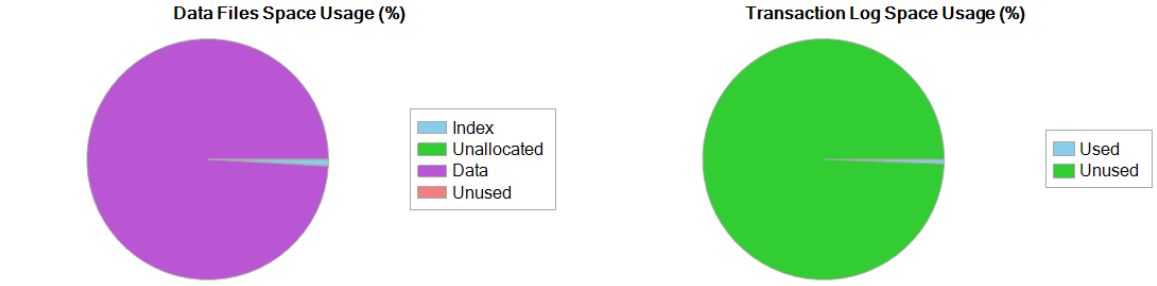
Não se tratando de um objetivo fundamental deste trabalho, a otimização do espaço ocupado não foi abordado. Para resolver este problema podem aplicar-se métodos de compressão dos dados a cada trajetória através da manutenção apenas dos registos de

Table Name	# Records	Reserved (KB)	Data (KB)	Indexes (KB)	Unused (KB)
dbo.dimCategoriaPontoEstadia	39	16	8	8	0
dbo.dimData	1.892	264	200	16	48
dbo.dimDispositivoCaptura	1	16	8	8	0
dbo.dimLocalizacao	256	264	232	16	16
dbo.dimPontePontoEstadia	5.733	136	104	16	16
dbo.dimPontoEstadia	4.912	1.736	1.712	16	8
dbo.dimTempo	86.400	3.592	3.576	16	0
dbo.dimTipoMovimento	12	16	8	8	0
dbo.dimTrajetoria	18.670	712	696	16	0
dbo.dimUtilizador	182	16	8	8	0
dbo.dimUtilizadorPerfil	179	48	32	16	0
dbo.factMovimento	17.704.084	2.688.328	2.666.544	21.416	368

Tabela 4.6: Utilização do espaço em disco pelas tabelas no DW.

pontos de estadia, ao invés de todos os pontos capturados com o dispositivo GPS, tal como segue a abordagem de Richter *et al.* [34]. Porém, esta abordagem, tal como outras presentes na literatura [38] implicam a perda de propriedades dos dados das trajetórias (tal como, os registo entre o inicio e o fim de uma trajetória), o que contraria um dos objetivos do presente trabalho.

Total Space Usage:	18.255,69	MB
Data Files Space Usage:	2.637,00	MB
Transaction Log Space Usage:	15.618,69	MB



Disk Space Used by Data Files

Filegroup Name	Logical File Name	Physical File Name	Space Reserved	Space Used
PRIMARY	DWEspacioTemporal	C:\Program Files\Microsoft SQL Server\MSSQL10_50.MSSQLSERVER\MSSQL\DATA\DWEspacioTemporal.mdf	2,58 GB	2,57 GB

Tabela 4.7: Utilização do espaço em disco do DW.

4.3 Implementação do Cubo de Dados

A construção do cubo de dados foi elaborada utilizando o módulo *Analysis Services* [28] do SQL Server. Após a definição da fonte de dados, a elaboração do cubo é feita através da definição das dimensões, sendo neste processo definidas as hierarquias modeladas para efetuar técnicas como *drill-down* e *roll-up*. A construção do cubo é concluída com base na

tabela de factos, uma vez que todas as dimensões estão referenciadas nesta tabela através de chaves estrangeiras. Na Figura D.2, do Anexo D, é possível observar o cubo de dados final, no ambiente da ferramenta Analysis Services.

Um dos aspetos mais importantes neste processo é definir o modo como são guardados os dados multidimensionais [28]. O modo MOLAP (*Multidimensional On Line Analytical Processing)* caracteriza-se por permitir um reduzido tempo de resposta a interrogações, mas tem uma sincronização lenta com as fontes de dados. No modo ROLAP (*Relational On Line Analytical Processing*) apesar da sincronização ser mais rápida apresenta maiores tempos de resposta às interrogações. Finalmente, o modo HOLAP (*Hybrid On Line Analytical Processing*) é um sistema híbrido entre o MOLAP e ROLAP. Neste trabalho optou-se pelo modo MOLAP, pois embora torne o processamento do cubo mais moroso, os ganhos de desempenho nas interrogações são compensadores. Este fator é de extrema importância quando é utilizada a técnica *pivot* para visualização dos dados sobre diversas perspetivas e para escolha das dimensões pretendidas [19]

A integração da dimensão *Categoria Ponto de Estadia* foi realizada através da abordagem de Ferrari *et al.* [15], que utiliza a dimensão de ponte definida anteriormente. Esta dimensão de ponte é definida como um *measure group*, interligando assim a dimensão de *Ponto de Estadia* e as respetivas categorias.

O problema da contagem distinta é nesta fase aperfeiçoado através da criação de medidas [40] que visam a contagem distinta de colunas da tabela de factos. Assim foram criadas medidas de contagem distinta para as chaves da dimensões *Trajetória*, *Ponto de Estadia*, *Localização* e *Utilizador* de modo a identificar univocamente cada registo destas dimensões.

Adicionalmente, foram ainda criadas as seguintes medidas calculadas [40]: *TempoDecorridoMinutos* (*TempoDecorrido* / 60), *DistanciaPercorridaMedia* (*DistanciaPercorrida* / Numero de registos), e *VelocidadeMedia* (*Velocidade* / Numero de registos).

4.4 Experimentação do Modelo

Após a concretização do modelo e do respetivo cubo de dados, esta secção apresenta a experimentação do modelo de forma a exercitar os níveis de análise e flexibilidade do mesmo. A metodologia seguida, baseia-se no conjunto de interrogações nos processos de negócio definidos na Secção 3.1. Na Secção 4.4.2 é apresentada uma aplicação de visualização de trajetórias desenvolvida para demonstrar a utilidade do modelo na integração com aplicações SIG. Por fim, é realizada uma pequena demonstração das interrogações através dos resultados obtidos do método de descoberta de utilizadores semelhantes.

4.4.1 Exemplos de Uso com Interrogações Analíticas

Nesta secção pretende-se validar o modelo concretizado através dos resultados de um conjunto de interrogações formuladas a partir das questões apresentadas na Secção 3.1. Estas interrogações para além de estarem relacionadas com os processos de negócio abrangidos pelo modelo, pretendem demonstrar as sua capacidades para responder a interrogações que exijam a aplicação de técnicas OLAP (Secção 2.1.3). Os resultados destas interrogações são apresentados de acordo com a técnica *pivot* através da ferramenta Microsoft Excel, mais concretamente *Excel Pivot Tables* que permitem a ligação com o cubo de dados (ou servidor OLAP) criado no Analysis Service.

De seguida serão discutidas as interrogações e respetivos resultados.

- **Qual é o total de utilizadores que se movimentam no bairro mais visitado de Beijing num dado intervalo de tempo?**

  Antes de efetuar esta interrogação foi verificado (através do número de presenças distintas relativas a utilizadores) que o bairro mais movimentado de Beijing era Zhongguancun. Para realizar a interrogação proposta foi então necessário aplicar diversas técnicas OLAP: na hierarquia da dimensão *Tempo* foi efetuado um *drill-down* para o segundo nível (*PeriodoDia* $>$ *Hora*) e aplicada a operação *slice* para obter um período temporal entre as 8h e as 21h. A mesma operação foi usada para selecionar o bairro de Zhongguancun da dimensão *Ponto de Estadia*, e por fim foi utilizada a técnica de contagem distinta para obter os diferentes utilizadores que frequentaram o bairro ao invés da contagem de todas as visitas desses mesmos utilizadores no período estipulado.

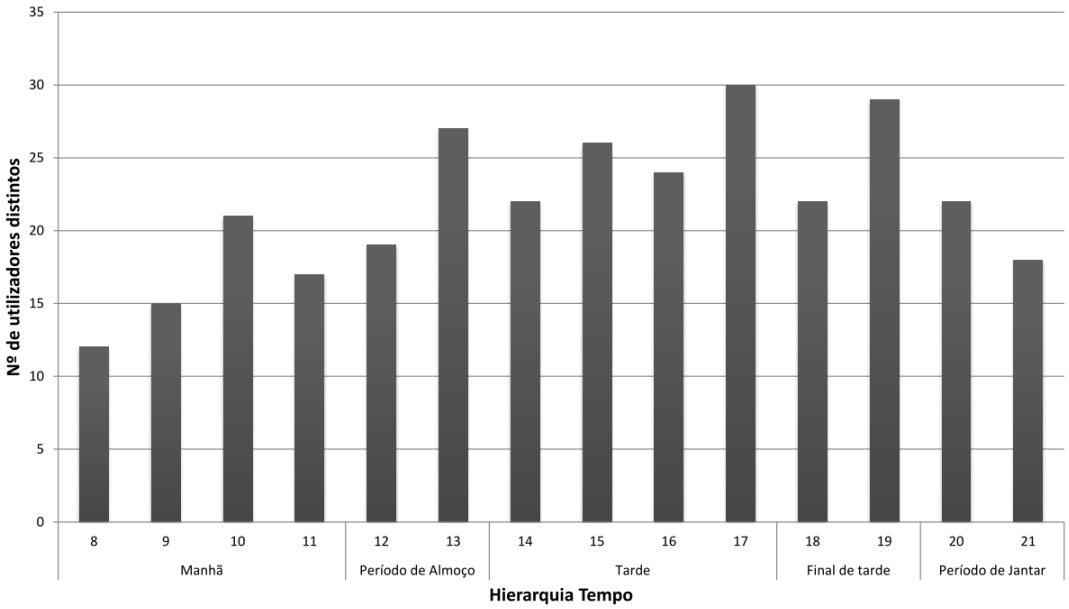
  No gráfico da Figura 4.6(b) é possível observar que este bairro é potencialmente uma zona laboral, pois o maior número de utilizadores ocorre entre as 10h e as 19h (ver também a tabela da Figura 4.6(a), em que este número nas restantes horas do dia decresce (inclusive nas horas não consideradas da análise).

- **Existe alguma diferença substancial na velocidade média dos veículos em Beijing durante o fim-de-semana?**

  Nesta interrogação foi utilizada a operação *slice* para selecionar apenas os tipos de movimento carro, mota, táxi e autocarro. Para isto foi necessário o atributo *Dia da Semana Textual* e a hierarquia da dimensão *Tempo*, tal como a medida agregada *VelocidadeMedia* (valor em metros/segundo). Pela análise da Tabela 4.8, é possível observar que ao fim de semana a velocidade a que os veículos circulam em Beijing aumenta ligeiramente, sendo o aumento mais significativo ao domingo. De realçar que este tipo de análises são tipicamente de grande importância para instituições

Bairro	Zhongguancun
Row Labels	**Utilizador Distinct Count**
Manhã	**30**
8	12
9	15
10	21
11	17
Período de Almoço	**31**
12	19
13	27
Tarde	**52**
14	22
15	26
16	24
17	30
Final de tarde	**37**
18	22
19	29
Período de Jantar	**29**
20	22
21	18
Grand Total	**66**

(a)



(b)

Figura 4.6: Presença de utilizadores no bairro Zhongguancun por períodos do dia em tabela (a) e em gráfico (b).

de planeamento urbano, tais como câmaras municipais, segurança pública, entre outras.

- **Quais os transportes onde os utilizadores passam mais tempo, por períodos do dia, durante a semana e fim de semana?**

  Com esta interrogação pretende-se analisar os hábitos dos utilizadores em relação à utilização de transportes públicos. Para isso, foi utilizado o atributo *coletivo* da di-

Nome	(Multiple Items)
Row Labels	**VelocidadeMedia**
Segunda-Feira	**6,07**
Terca-Feira	**6,33**
Quarta-Feira	**6,43**
Quinta-Feira	**6,24**
Sexta-Feira	**6,62**
Sabado	**6,96**
Domingo	**7,33**
Madrugada	7,07
Manhã	8,23
Período de Almoço	7,65
Tarde	6,55
Final de tarde	7,77
Período de Jantar	7,31
Noite	7,44
Grand Total	**6,63**

Tabela 4.8: Velocidade média de veículos por dia da semana.

mensão *Tipo de Movimento*, sendo o tempo decorrido em minutos distribuído pelo atributo *DiaSemanaTextual* e pela hierarquia da dimensão*Tempo*. Pode-se observar na Tabela 4.9 que os transportes mais utilizados são o autocarro e o comboio, não existindo diferenças significativas entre o metro e o táxi. Quanto ao tipo de movimento por avião, embora tenha um valor considerável, possui poucos registos para ser relevantes.

- **Nos bairros de Beijing quais os períodos do dia em que existe mais movimentação de veículos e a velocidade média de circulação dos mesmos?**

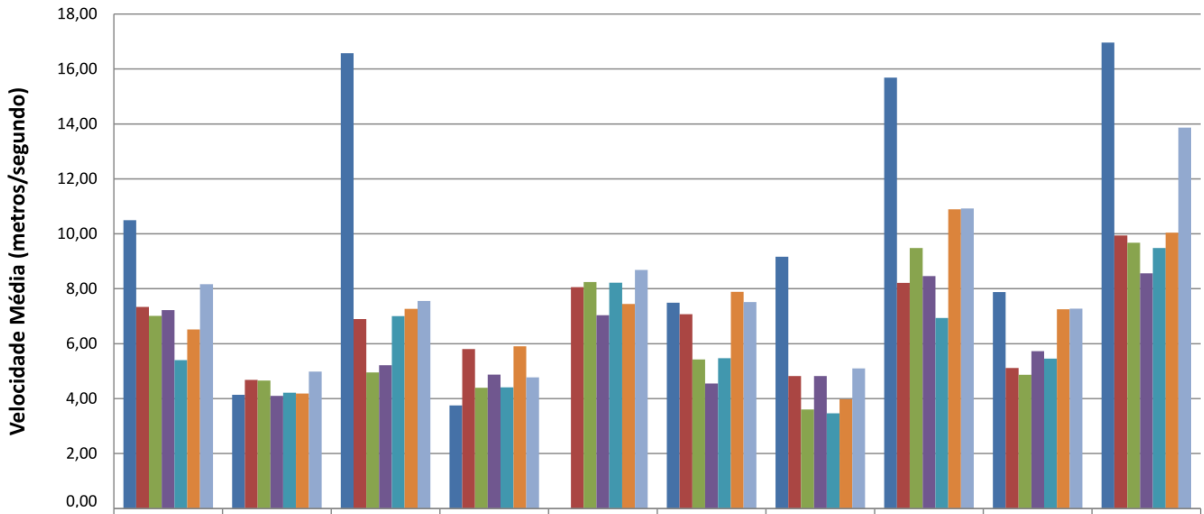
  Esta interrogação está intrinsecamente ligada com processo de planeamento de tráfego urbano. Perceber quais são as localizações da cidade com mais movimentação de veículos e a velocidade a que estes se deslocam tem uma grande utilidade para as autoridades responsáveis pelo controlo de tráfego urbano. Para responder a esta questão foram selecionadas as localizações com mais movimento de utilizadores distintos (contagem distinta por utilizador) em que o tipo de movimento era o carro, a mota, o táxi ou o autocarro. A medida utilizada foi *VelocidadeMedia* vista pelas dimensão de *Localização* e hierarquia da dimensão *Tempo*.

  É possível observar no gráfico da Figura 4.7 que existem localizações onde a velocidade é bastante elevada de madrugada (por exemplo, localizações 53 e 165), o que poderá significar que devem tratar-se de vias bastante utilizadas pelos utilizadores quando vão para o emprego. É também possível verificar que as localizações 135 e 165 têm sempre médias bastante elevadas, sendo possível que correspondam a autoestradas de Beijing.

Transporte Coletivo	Verdadeiro

TempoDecorridoMinutos	Column Labels					
Row Labels	Autocarro	Avião	Barco	Comboio	Metro	Taxi
Segunda-Feira	**9549**	**24**		**18203**	**2835**	**1518**
Terca-Feira	**76337**	**122**		**9804**	**3447**	**851**
Madrugada	341					15
Manhã	68142			774	1517	173
Período de Almoço	828	122			43	21
Tarde	3410			8921	596	156
Final de tarde	2071			41	307	220
Período de Jantar	1253			66	287	160
Noite	292			2	698	108
Quarta-Feira	**10340**	**3360**		**3156**	**4247**	**1671**
Quinta-Feira	**10266**	**28**		**2435**	**3045**	**1411**
Sexta-Feira	**19599**	**7020**		**13593**	**2378**	**4749**
Sabado	**33466**		**64**	**17602**	**3882**	**3543**
Madrugada	2000			12		41
Manhã	2693			878	758	611
Período de Almoço	13051			15708	473	332
Tarde	4297		64	203	1022	593
Final de tarde	9693				402	1179
Período de Jantar	1275			800	937	402
Noite	457				289	386
Domingo	**21444**	**3138**		**19540**	**2994**	**2419**
Grand Total	**181001**	**13692**	**64**	**84333**	**22828**	**16163**

Tabela 4.9: Distribuição de tempo passado em transportes.



Localizações	8	44	53	56	68	69	73	135	137	165
Madrugada	10,50	4,14	16,57	3,75		7,50	9,17	15,68	7,88	16,96
Manhã	7,34	4,68	6,89	5,81	8,06	7,08	4,82	8,21	5,12	9,94
Período de Almoço	7,01	4,65	4,95	4,40	8,25	5,43	3,61	9,49	4,87	9,68
Tarde	7,22	4,10	5,22	4,87	7,03	4,55	4,82	8,46	5,73	8,56
Final de tarde	5,40	4,22	7,00	4,41	8,22	5,47	3,46	6,94	5,45	9,49
Período de Jantar	6,52	4,19	7,27	5,90	7,44	7,89	3,97	10,89	7,25	10,04
Noite	8,16	4,98	7,55	4,78	8,69	7,52	5,10	10,92	7,28	13,86

Figura 4.7: Localizações com mais movimentação e velocidade média.

- **Quais as horas e períodos do dia em que existe mais e menos movimentação (dinâmica da cidade)?**

  Para entender os momentos do dia em que Beijing tem mais movimentação, foram considerados dois momentos distintos: dias úteis e fim de semana (excluindo feriados). Foram utilizadas as medidas *DistanciaPercorridaMedia* e *VelocidadeMedia* distribuídas por dias úteis e fim de semana (utilizando o atributo *Feriado* para excluir os mesmos da análise do fim de semana), e por períodos do dia (*HierarquiaTempo*). Através da manipulação de atributos, esta análise poderia ser feita para analisar a movimentação por diferentes períodos, tal como, mensal ou anual.

  Como esperado é possível observar, pelo gráfico na Figura 4.8 que os períodos de maior movimentação nos dias úteis ocorrem em momentos diferentes dos do fim de semana, sendo o pico de movimentação mais cedo nos dias úteis. Ao fim de semana existe um ligeiro aumento de movimentação durante o início da madrugada e período da tarde, provavelmente derivado às saídas noturnas e a saídas para passear, respetivamente.

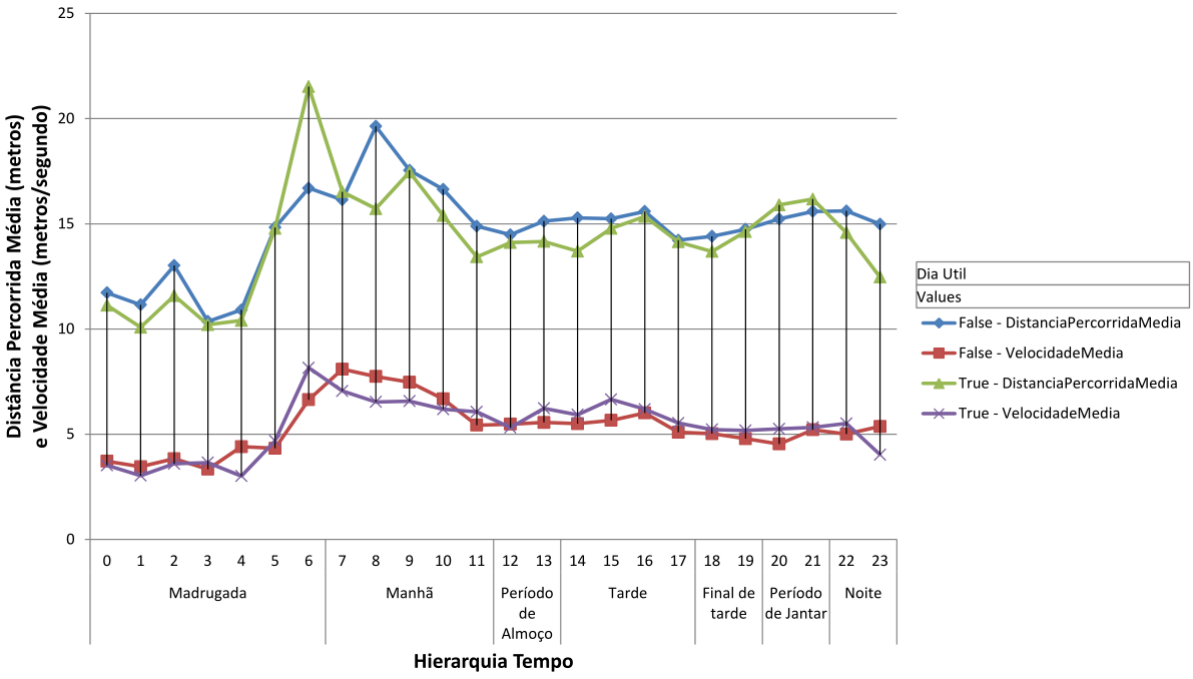
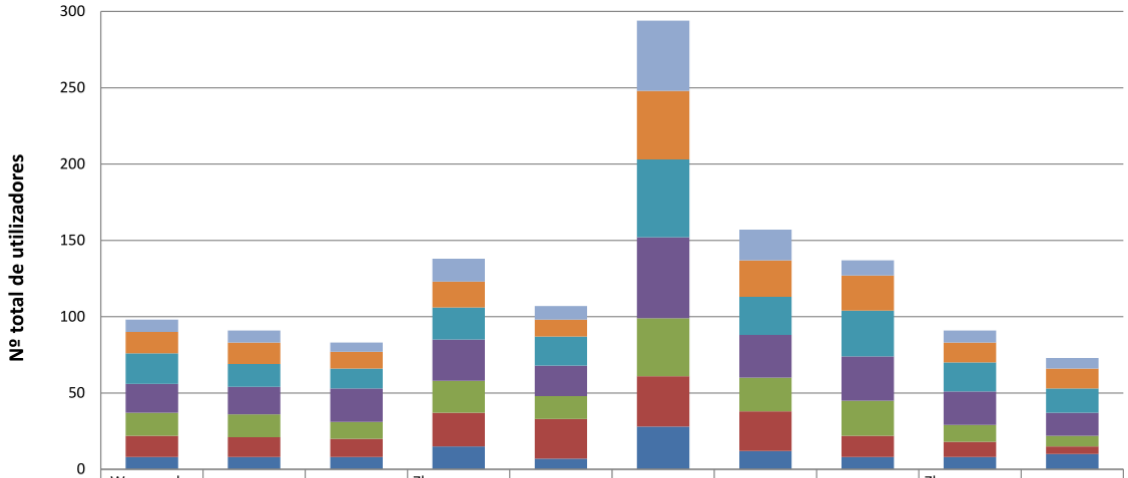


Figura 4.8: Dinâmica da vida em Beijing.

- **Para cada período do dia e dia da semana, qual a distribuição de visitas por diferentes utilizadores por ponto de estadia mais visitado?**

  Esta interrogação parte do pressuposto que os pontos de estadia mais visitados são aqueles que são mais visitados por diferentes utilizadores, ao contrário de ser muito visitado mas por apenas um número limitado de utilizadores. Foi então utilizada a

contagem distinta de utilizadores distribuída pelo atributo *Rua* da dimensão *Ponto de Estadia* e pelos períodos do dia utilizando a hierarquia da dimensão *Tempo*. No gráfico da Figura 4.9 pode-se observar que o ponto de estadia mais visitado se encontra na Zhichun Road tendo um número de visitas de utilizadores bem distribuída por diferentes períodos do dia.



	Wanquanhe Road Side Road	Suzhou Street	Zhichun Road	Zhongguancun North Street	Beiqingyun Road	Zhichun Road	Zhongguancun East Road	Chengfu Road	Zhongguancun North 2nd Street	Caihefang Road
Noite	8	8	6	15	9	46	20	10	8	7
Período de Jantar	14	14	11	17	11	45	24	23	13	13
Final de tarde	20	15	13	21	19	51	25	30	19	16
Tarde	19	18	22	27	20	53	28	29	22	15
Período de Almoço	15	15	11	21	15	38	22	23	11	7
Manhã	14	13	12	22	26	33	26	14	10	5
Madrugada	8	8	8	15	7	28	12	8	8	10

Figura 4.9: Ruas mais frequentadas de Beijing.

- **Quais as localizações de Beijing mais ativas, ou seja, com mais movimentação, em horário laboral aos dias de semana? E em horário pós-laboral?**

  Seguindo a abordagem de Becker *et al.* [6], para entender quais as localizações de uma cidade em que se concentram as áreas laborais e áreas de habitação, é necessário observar as localizações onde existe maior movimentação em horário laboral e pós-laboral. Para tal, foi efetuado um filtro (*slice*) por dias úteis, e utilizada a medida *DistanciaPercorridaMedia* distribuída por localizações. No gráfico da Figura 4.10 é possível observar a combinação das localizações mais movimentadas em horário laboral e em horário pós laboral. Tendo em conta a abordagem referida, pode-se concluir que as localizações 248 e 253 serão áreas residenciais, sendo a localização 20, 250 e 252 aquelas que se destingem como áreas laborais.

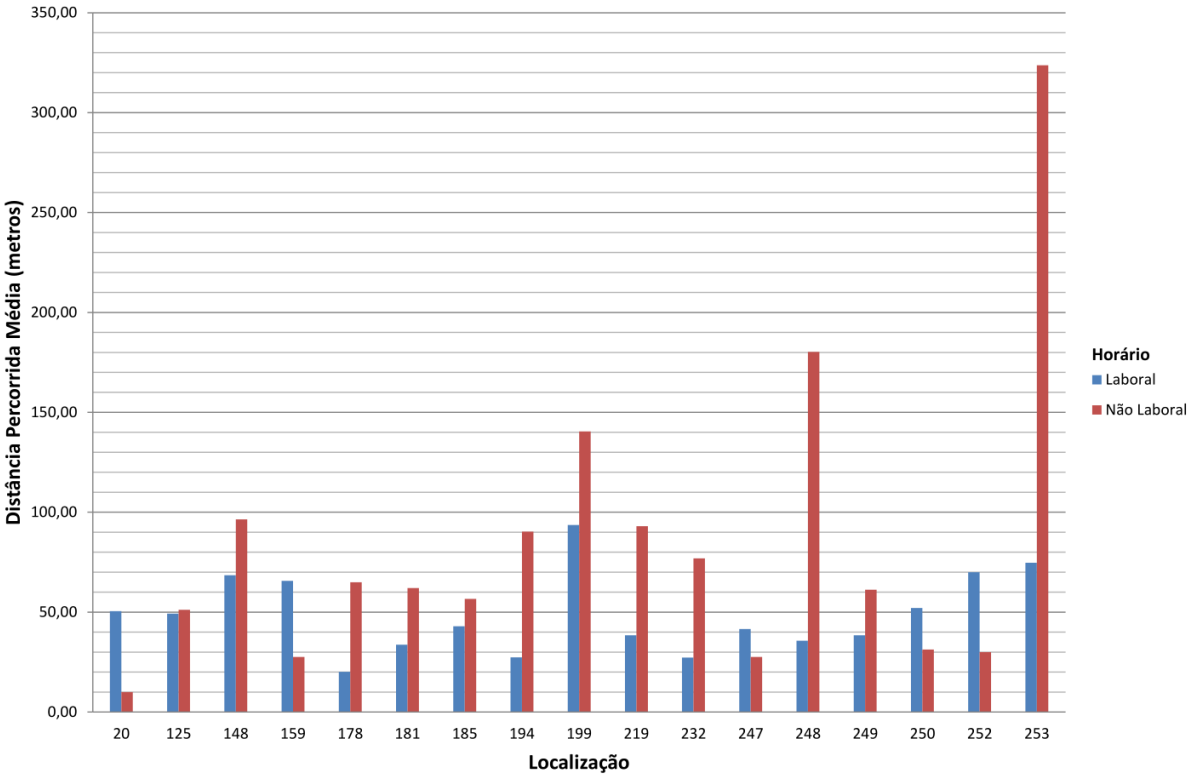


Figura 4.10: Localizações laborais e habitacionais.

- **Quais os pontos de estadia de divertimento noturno de Beijing com mais movimento?**

  Para realizar esta análise foi feita uma combinação do período das 22h às 2h de sexta e sábado, utilizando a hierarquia *Tempo* e o atributo *DiaSemana*. A medida calculada *TempoDecorridoMinutos* foi distribuída pelos pontos de estadia através do atributo *Rua*. No gráfico da Figura 4.11 pode-se observar que as ruas mais relevantes são: Qinghua West Road, Zhincun Road e Zhongguancun North Street.

- **Quais os pontos de estadia mais visitados e a duração do tempo de estadia dos utilizadores durante um determinado período de festividades na cidade?**

  Após a realização de eventos de grande dimensão, existem diversas análises que podem ser feitas, como perceber o que mudou no tráfego urbano ou se existiu um grande impacto na frequência de utilizadores nas localizações onde se realizaram os eventos. O conjunto de dados Geolife abrange o período temporal dos jogos olímpicos realizados em Beijing em 2008[12]. Para perceber os pontos de estadia mais visitados neste período, foi selecionado o período temporal de 08/08/2008 a 24/08/2008, sendo utilizada a medida calculada *TempoDecorridoMinutos* distribuída pelo atributo *Rua* da dimensão *Ponto de Estadia* e hierarquia da dimensão

[12]http://en.beijing2008.cn/

Tempo. Os resultados desta interrogação (gráfico da Figura 4.12) pretendem apenas demonstrar a utilidade analítica do modelo também em períodos em que a dinâmica normal de uma região se altere.

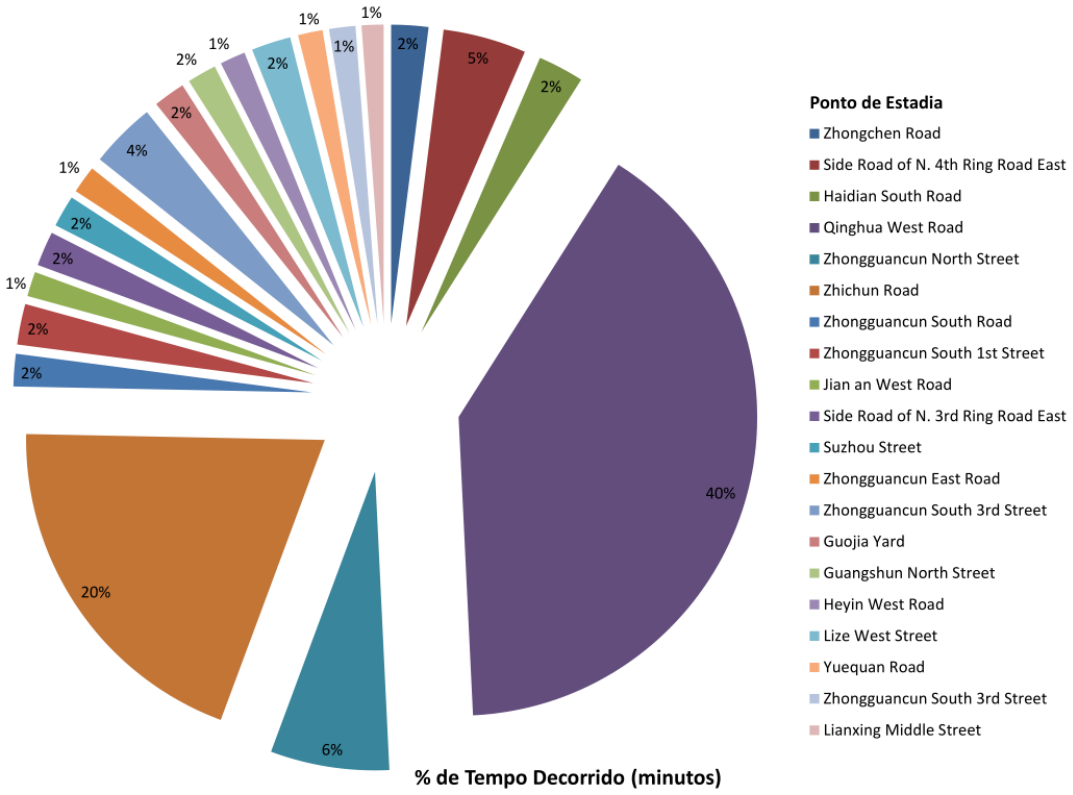


Figura 4.11: Pontos de estadia noturnos mais frequentados.

Através das interrogações realizadas, pretende-se demonstrar a utilidade do modelo nos processos de negócio mencionados, em particular na análise de planeamento urbano. Foram também apresentadas análises que potencializassem o uso de técnicas OLAP, tendo sido usadas maioritariamente as técnicas *drill-down*, *slice* e *dice*, para além da técnica *pivot* que é base da análise multidimensional.

Embora se tenha analisado o conjunto de dados Geolife sob diversas perspetivas, um dos problemas associados com os resultados apresentados está relacionado com o facto de este não possuir informações sobre o perfil dos utilizadores. Este facto diminui a quantidade de interrogações possíveis, que certamente permitiriam analisar os dados sob outra perspetiva. Por exemplo, interrogações como 'qual o intervalo de idades dos utilizadores que frequentam o ponto de estadia x?' ou 'existe diferença na movimentação diária de utilizadores dos diferentes sexos?', poderiam ser respondidas pelo DW concretizado, caso se tivesse informação das idades dos utlizadores.

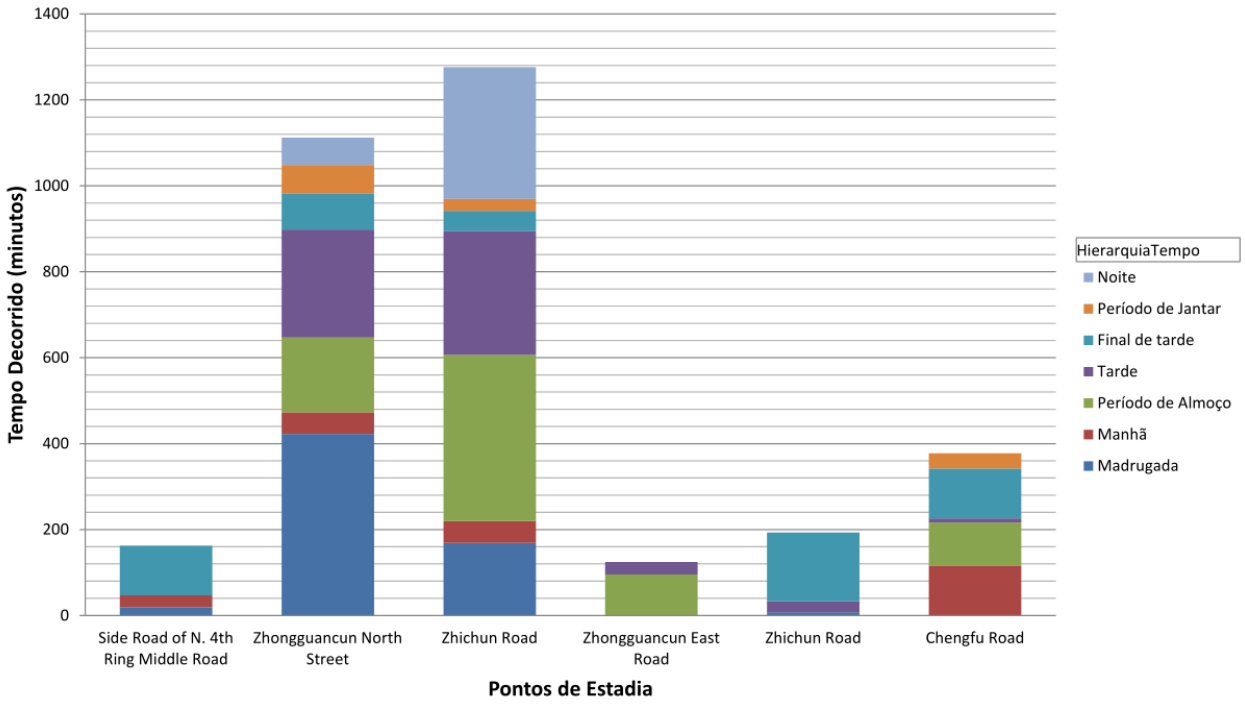


Figura 4.12: Pontos de estadia mais frequentados durante os jogos olimpicos de Beijing.

4.4.2 Visualização dos Dados

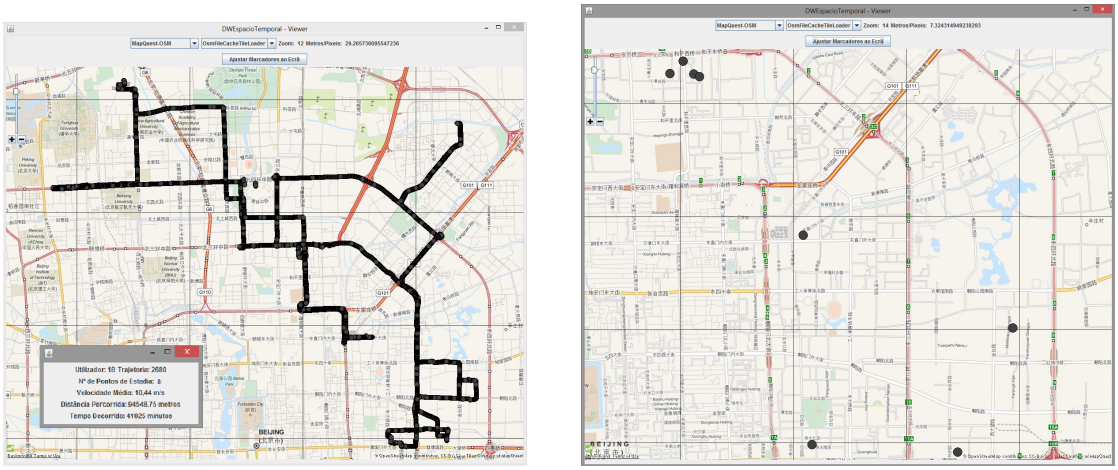
O modelo proposto neste projeto tem como foco a análise multidimensional de informação espácio-temporal. Porém, quando se manipula este tipo de informação é importante que seja possível visualizá-la através da sua representação no espaço. Para conseguir isso é necessário que exista uma integração do DW com aplicações de visualização ou SIGs. Com o modelo apresentado é possível efetuar essa integração pois é guardada a informação geográfica dos registos espácio-temporais, ao contrário do que sucede com alguns dos trabalhos apresentados na Secção 2.1.4. Para demonstrar visualmente os dados presentes no DW foi desenvolvida uma aplicação em Java (adaptada do JMapViewer[13]). Adicionalmente é possível visualizar os pontos espaciais das trajetórias, localizações e pontos de estadia associados aos utilizadores.

Na Figura 4.13(a) pode-se observar uma trajetória do utilizador 18 e o número de pontos de estadia respetivos podem ser observados na Figura 4.13(b).

Na Figura 4.14 podemos visualizar todos os pontos de estadia na cidade de Beijing que foram extraídos do conjunto de dados através do método apresentado na Secção 3.2.3.

Por fim, na Figura 4.15 pode-se observar as localizações de *Nível 2* que foram criados com o método de agrupamento de localizações, explicado na Secção 3.2.2.

[13]http://wiki.openstreetmap.org/wiki/JMapViewer



(a) (b)

Figura 4.13: (a) Representação visual de uma trajetória e (b) respetivos pontos de estadia.

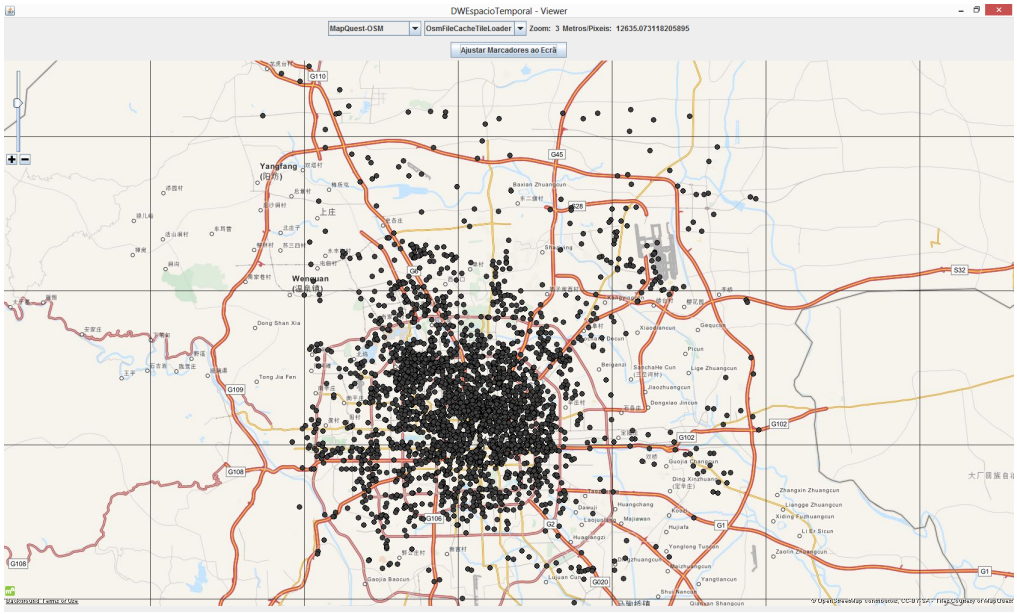


Figura 4.14: Pontos de estadia extraídos das trajetórias do conjunto de dados Geolife.

4.4.3 Utilizadores Semelhantes

Nesta secção são apresentados alguns exemplos de interrogações OLAP e de visualização geográfica de informação sobre os grupos de utilizadores criados através do método apresentado na Secção 3.2.4.

- **Quais os pontos de estadia e suas categorias que o grupo 1 de utilizadores semelhantes costumam visitar por períodos do dia?**

  Para calcular os pontos de estadia mais frequentados por utilizadores do grupo 1, foram selecionados apenas os utilizadores com o valor de '1' no atributo *Clus-*

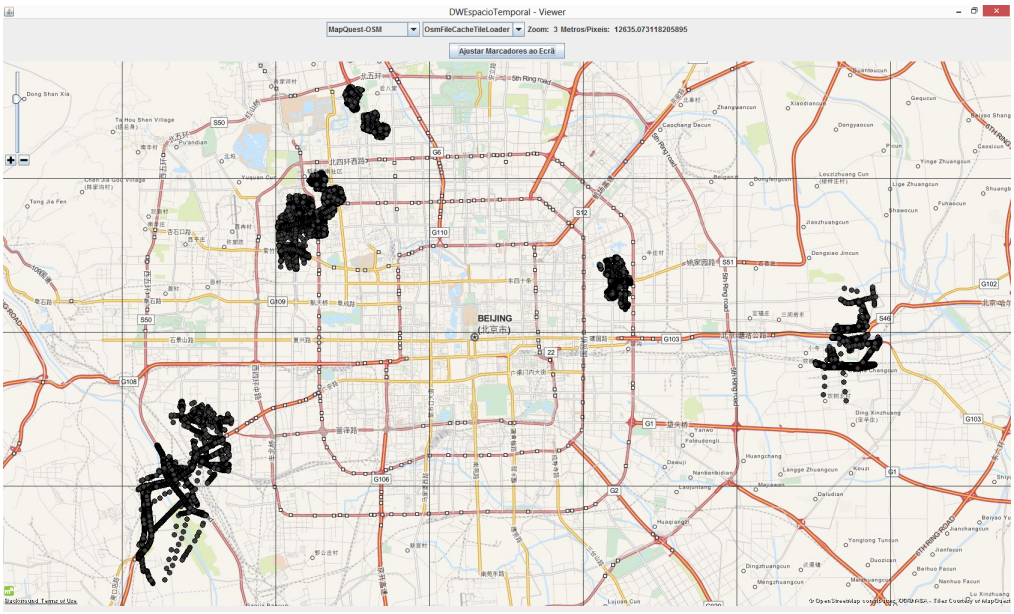


Figura 4.15: Representação das localizações de *Nível 2* da dimensão *Localização*.

ter, sendo selecionado os atributos *Rua* e *Nome* da dimensão *Ponto de Estadia* e *Categoria* da dimensão *Categoria do Ponto de Estadia*. A medida *TempoDecorridoMinutos* foi distribuída pela hierarquia da dimensão *Tempo*, utilizando a medida de contagem distinta de utilizadores para identificar os pontos de estadia mais frequentados.

No gráfico da Figura 4.16 pode-se observar que o ponto de estadia com mais relevância para os utilizadores do grupo 1, está localizado na rua Zhichun Road No. 82 e é designado no Google Places como 'estabelecimento', sendo o ponto de estadia mais semelhante entre estes utilizadores.

- **Quais os pontos de estadia da categoria 'universidade' que os utilizadores semelhantes do grupo 3 costumam frequentar, por períodos do dia?**

  Atualmente, as campanhas de *marketing* para jovens são cada vez mais realizadas em locais de ensino. Para responder à questão apresentada, foram selecionados os utilizadores com o valor '3' no atributo *Cluster*, utilizado o atributo *Categoria* com o valor 'universidade' para filtrar os pontos de estadia pela categoria pretendida. A medida *TempoDecorridoMinutos* foi distribuída pela hierarquia da dimensão *Tempo*, obtendo-se o gráfico apresentado na Figura 4.17. Este gráfico permite concluir que a universidade *Chinese Academy of Sciences Kindergarten* é ideal para uma campanha direcionada a estudantes (assumindo que os utilizadores são estudantes), e que mediante a análise do tempo despendido pelos utilizadores deve ser feita no período da tarde e final da tarde.

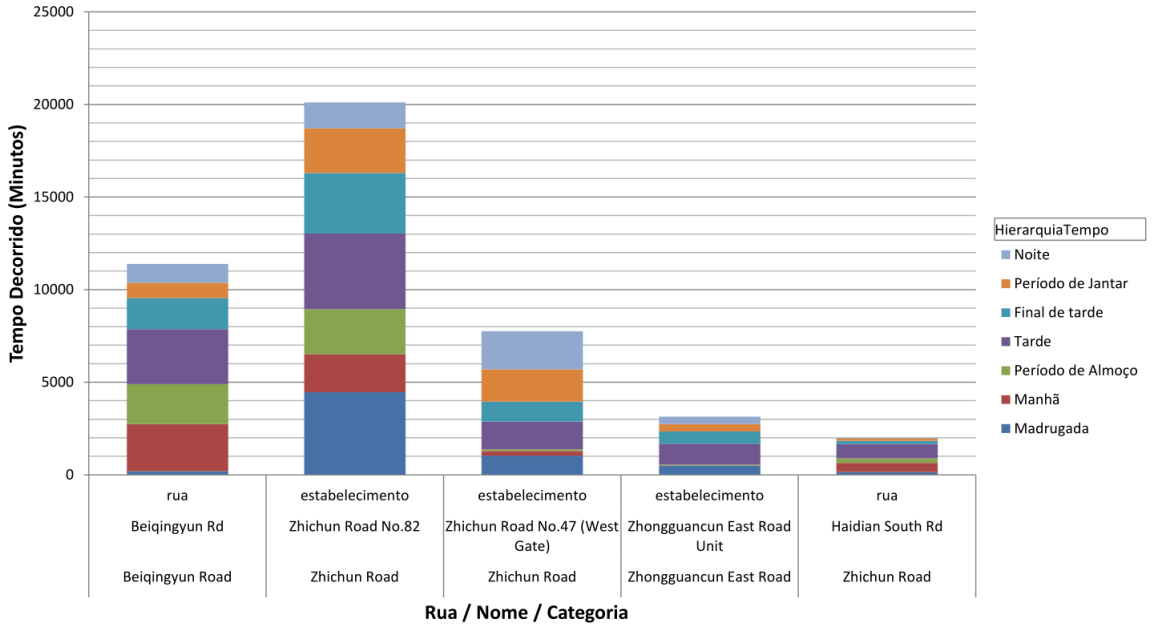


Figura 4.16: Pontos de estadia frequentados por utilizadores do grupo 1.

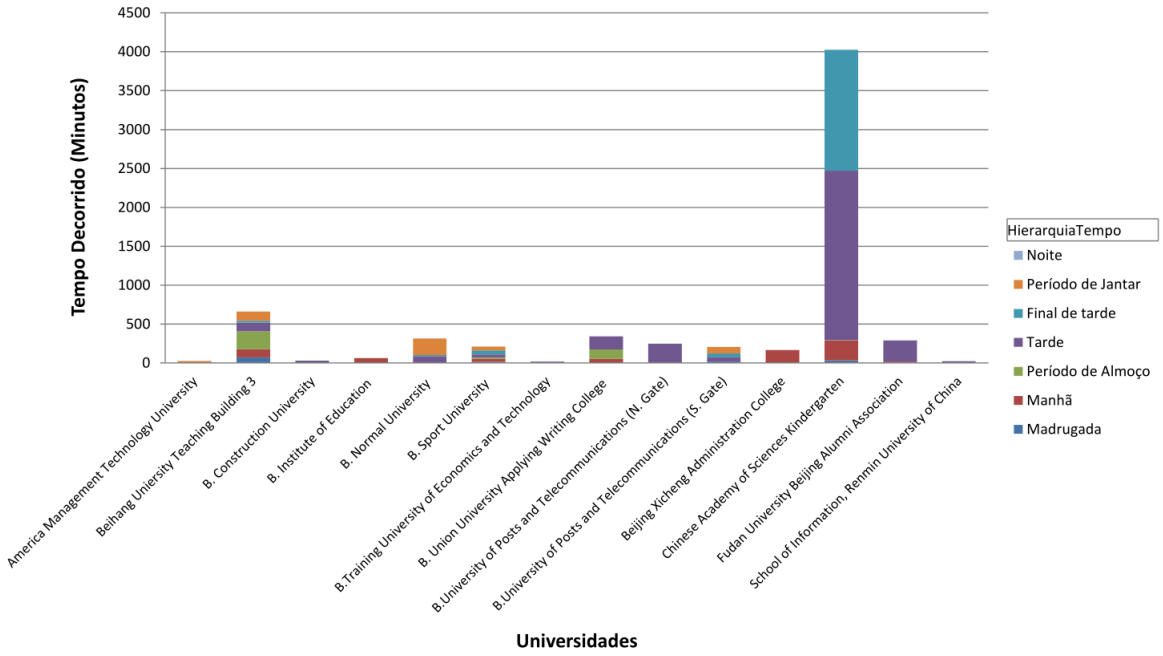


Figura 4.17: Universidades frequentadas por utilizadores do grupo 3.

- **Distribuição geográfica dos pontos de estadia frequentados pelos utilizadores semelhantes do grupo 3.**

  Na Figura 4.18 estão representados os pontos de estadia que são frequentados pelos utilizadores do grupo 3. Pode-se observar que a maioria dos pontos de estadia se encontram localizados na mesma região, existindo apenas alguns mais afastados.

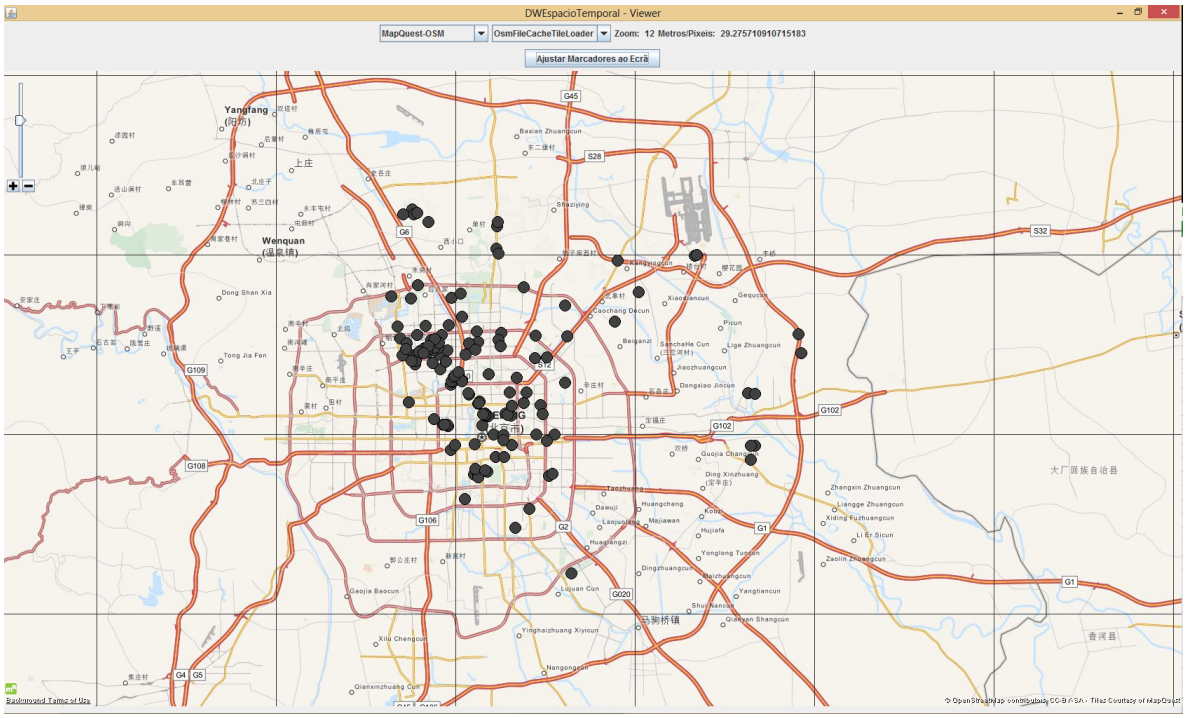


Figura 4.18: Pontos de estadia frequentados por utilizadores do grupo 3.

Nesta secção apresentou-se um possível conjunto de interrogações com o conjunto de dados Geolife. Contudo a ausência de dados demográficos dos utilizadores não permite aprofundar outras questões, como por exemplo, 'qual é faixa etária do ponto de estadia mais frequentado pelos utilizadores pertencentes ao grupo 1?'. A experimentação do modelo com um conjunto de dados mais rico semanticamente do ponto de vista demográfico irá ser um dos objetivos para trabalho futuro.

Capítulo 5

Conclusão

Os avanços recentes nas tecnologias de posicionamento e a melhoria das características dos telemóveis e *smartphones*, nomeadamente, a disponibilização frequente de dispositivos de GPS, têm reforçado a capacidade de recolha de grandes quantidades de dados de posicionamento sobre as pessoas, veículos, ou outros objetos em movimento. O crescente interesse por esta informação está relacionado com a sua aplicação na área dos sistemas de recomendação baseados na localização e no contexto (por exemplo: recomendação de restaurantes dependente do local em que o utilizador se encontre: trabalho ou perto de casa). Uma das áreas que tem sido alvo de estudo com o objetivo de organizar, gerir e extrair informação relevante, tem sido o estudo de comportamentos humanos através das suas trajetórias. De modo a extrair informação útil e relevante é fundamental a conceção de métodos adequados para o tratamento, análise, descoberta de conhecimento e prospeção de dados. Contudo, os dados existentes sobre a mobilidade humana apresentam ainda redundâncias, incoerências, pouca informação semântica e ainda são escassos os modelos de dados e algoritmos de prospeção de dados especialmente concebidos para este tipo de dados, espácio-temporais.

Nesta dissertação foi apresentada a conceção de um *Data Warehouse* Espácio-Temporal orientado para análise de dados de trajetórias de utilizadores. O modelo proposto apresenta uma caracterização semântica dos registos, quer a nível de utilizadores e movimento dos mesmos, geográfico, temporal, quer ainda a nível técnico de captura dos registos. Foi ainda desenvolvido no processo ETL, algoritmos para caracterização de trajetórias, extração de pontos de estadia e a sua categorização e divisão de localizações da região abrangida.

Para suprimir o problema da pouca informação apresentada pelos dados sobre trajetórias, foram criados métodos de enriquecimento semântico de trajetórias para integrar o processo ETL. Para o enriquecimento de características numéricas de trajetórias, foram utilizadas diversas fórmulas matemáticas com base nos dados espaciais e temporais de

cada registo, de forma a calcular informações como a distância percorrida, velocidade média do movimento, e tempo decorrido desde o anterior registo.

A criação de conjuntos de localizações foi feita através do agrupamento de registos geográficos pertencentes às trajetórias dos utilizadores móveis. Este agrupamento foi efetuado através de um método hierárquico aglomerativo para efetuar a criação de grupos de registos que estivessem próximos geograficamente. Para abordar o problema de extração de pontos de estadia de uma trajetória, foram identificadas as situações em que estes pontos ocorriam na mesma, tendo sido elaborado um algoritmo de extração de pontos de estadia com base na distância espacial e temporal entre pontos. Foi ainda realizada a caracterização destes pontos de estadia, através de serviços de obtenção de informação geográfica.

Com o objetivo de descobrir utilizadores que tivessem os mesmo hábitos de movimentação, foi criado um método que permite agrupar utilizadores com base nos seus pontos de estadia e localizações frequentadas. O processo desenvolvido além de identificar os pontos de estadia e localizações em comum entre utilizadores, atribui-lhes um peso que expressa a importância de cada ponto de estadia e localização para cada utilizador. A razão deste processo ter sido criado apenas tendo por base informação geográfica, deve-se ao facto da maioria dos conjuntos de dados de trajetórias não conterem informação de perfil dos utilizadores.

A concretização do modelo foi efetuada através do conjunto de dados Geolife da *Microsoft Reasearch Asia* composto por 182 utilizadores abrangendo um período temporal de 5 anos, contendo no total 18 670 trajetórias. Através do processo ETL o conjunto de dados foi reduzido para apenas conterem os dados referentes à cidade Beijing, foi efetuado tratamento destes dados, eliminando duplicados, *outliers*, entre outros. Através da aplicação dos métodos criados, foram gerados cerca de 4 000 pontos de estadia (devidamente caracterizados), tal como 256 grupos de localizações. Após o carregamento dos dados, foi implementado um cubo de dados multidimensional para realizar análises que permitam extrair informação relevante, tal como a utilização de hierarquias e técnicas OLAP.

A experimentação do modelo foi efetuada tendo por base interrogações que foram definidas na listagem de prioridades para o modelo dimensional. As análises efetuadas foram direcionadas para as áreas de planeamento urbano, gestão de tráfego, assim como para a análise de movimentação dos utilizadores, permitindo os resultados obtidos validar a utilidade do modelo proposto para essas áreas de negócio. Foi ainda desenvolvida uma aplicação para visualização dos dados presentes no DW com o objetivo de demonstrar a viabilidade de interligação do DW com aplicações SIG e de visualização de dados geográficos.

Como trabalho futuro, existem diversos pontos em que o modelo proposto pode ser melhorado:

- De forma a estruturar melhor a informação geográfica, poderia explorar-se uma representação de grelha regular, adotando assim uma representação bastante utilizada na área do planeamento urbano, realizando ao mesmo tempo a resolução do relacionamento parcialmente contido do qual esta abordagem aderece;
- Criação de um método de reconstrução de trajetórias, para a utilização de conjuntos de dados em bruto de movimentação de objetos móveis, que permita assim extrair as trajetórias e pontos de estadia que fazem parte da mesma;
- Refinamento do processo de descoberta de utilizadores semelhantes através da inclusão do fator temporal de presença nos pontos de estadia e localizações analisadas, permitindo assim uma maior precisão na semelhança dos utilizadores. Outra abordagem para este processo seria a identificação de trajetórias semelhantes, ou seja, utilizadores que tivessem o mesmo tipo de movimentação (sem considerar pontos de estadia e/ou localizações);
- Por fim, o maior objetivo futuro será utilizar um outro conjunto de dados (por exemplo, referente à mobilidade de utilizadores nacionais), que permita obter mais informação a nível de pontos de estadia e tenha informações demográficas sobre os utilizadores, de modo a realizar-se análises de perfil mais enriquecidas, tal como a utilização de técnicas de prospeção de dados.

Apêndice A

Plano de Trabalhos

Tarefa	**Realizado**
Familiarização com o problema da exploração de trajetórias humanas e respetivas soluções na literatura	**1 Mês**
Modelação dimensional de um *data warehouse* espácio-temporal e técnicas de enriquecimento semântico de trajetórias	**2 Meses**
Concretização do *data warehouse* espácio-temporal	**6 Meses**
- Aplicação do modelo ao SQL Server	0,5 Meses
- Processo de extração e primeira fase do tratamento ao conjunto de dados	1 Mês
- Processo de aplicação das rotinas de enriquecimento semântico de trajetórias	1,5 Meses
- Carregamento dos dados e implementação do cubo de dados multidimensional	1 Mês
- Avaliação do modelo e realização de novas iterações ao processo de modelação/concretização	2 Meses
Escrita do artigo	**1 Mês**
Escrita da dissertação	**1 Mês**

Tabela A.1: Planeamento das atividades do projeto.

Apesar do projeto ter sido iniciado a 1 de Setembro de 2012, e ter como data final estimada 30 Junho de 2013, não foi possível entregar este documento na data prevista. Isto deveu-se a uma pequena reestruturação dos objetivos iniciais presentes no relatório preliminar, tal como à necessidade de participar em determinadas tarefas que não estavam inicialmente previstas, nomeadamente a escrita de um artigo sobre a conceção de um *data warehouse* espácio-temporal aplicado a trajetórias de utilizadores móveis:

- Vitor Oliveira, Ana Paula Afonso, André Falcão, *Conceção de um Data Warehouse Espácio-Temporal para Análise de Trajetórias Humanas*, no Simpósio de Informática INFórum, Évora, 2013.

Apêndice B

Data Warehouse - Tabelas

Este anexo contém as tabelas com a informação descriminada das dimensões e tabelas de factos criadas para o modelo proposto neste projeto.

Campo	Tipo de Dados	Descrição
IdData	Número Inteiro	Identificador único de registo
Dia	Número Inteiro	Identifica o dia do mês (valores de 1 – 31)
Mes	Número Inteiro	Identifica o mês do ano (valores textuais de 1 a 12)
Ano	Número Inteiro	Identifica o ano
DataCompleta	Data	Data completa no formato AAAA-MM-SS
DataCompletaTextual	Texto	Descrição textual do campo DataCompleta (ex: quinta-feira, 12 de Abril de 2007)
DiaSemana	Número Inteiro	Identifica o dia da semana (valores de 1 a 7)
DiaSemanaTextual	Texto	Descrição textual do campo DiaSemana (valores textuais de Segunda-Feira a Domingo)
DiaUtil	Número Inteiro	Identifica se o dia é um dia útil (1) ou não (0)
FimDeSemana	Número Inteiro	Identifica se é fim de semana (1) ou não (0)
Quinzena	Número Inteiro	Identifica a quinzena do mês (valores de 1 para primeira quinzena e 2 para segunda)
UltimoDiaMes	Número Inteiro	Identifica se o dia é o último dia do mês (1) ou não (0)
Feriado	Número Inteiro	Identifica se o dia é um feriado (1) ou não (0)
FeriadoTextual	Texto	Descrição textual do campo Feriado
MesTextual	Texto	Descrição textual do campo Mês (valores textuais de Janeiro a Dezembro)
EstacaoAno	Texto	Identifica a estação do ano

Tabela B.1: Representação detalhada da dimensão Data.

Campo	Tipo de Dados	Descrição
IdTempo	Número Inteiro	Identificador único de registo
Hora	Número Inteiro	Identifica a hora do dia (valores de 1 – 24)
Minuto	Número Inteiro	Identifica os minutos de uma hora (valores de 0-60)
Segundo	Número Inteiro	Identifica os segundos de um minuto (valores de 0-60)
HoraCompleta	Tempo	Hora completa no formato HH:MM:SS
PeriodoDia	Texto	Identifica o período do dia (Manhã 7h - 12h Período de Almoço 12h – 14h Tarde 14h – 18h Final de tarde 18h – 20h Período de Jantar 20h – 22h Noite 22h – 00h Madrugada 00h - 7h)

Tabela B.2: Representação detalhada da dimensão Tempo.

Campo	Tipo de Dados	Descrição
IdLocalizacao	Número Inteiro	Identificador único de registo
Nivel 1	Número Inteiro	Nivel 1 da Localização
Nivel 2	Número Inteiro	Nivel 2 da Localização
Nivel 3	Número Inteiro	Nivel 3 da Localização
Nivel n	Número Inteiro	Nivel n da Localização
Nivel 256	Número Inteiro	Nivel 256 da Localização
InicioValidade	Data	Início da validade do registo
FimValidade	Data	Fim da validade do registo
Razao	Texto	Razão de fim da validade do registo
EmVigor	Número inteiro	Se o registo está em vigor (0) ou não (1)

Tabela B.3: Representação detalhada da dimensão Localização.

Campo	Tipo de Dados	Descrição
IdUtilizador	Número Inteiro	Identificador único de registo
CodigoUtilizador	Número Inteiro	Identificador único do utilizador
Cluster	Número Inteiro	Identificador do grupo do utilizador
Sexo	Texto	Género do utilizador
Idade	Número Inteiro	Idade do utilizador
EstadoCivil	Texto	Estado civil do utilizador
HabilitacoesLiterarias	Texto	Habilitações literárias do utilizador
Nacionalidade	Texto	Nacionalidade do utilizador
Profissao	Texto	Profissão exercida pelo utilizador
Ativo	Número Inteiro	Se o utilizador está em ativo profissionalmente (1) ou não (0)
ClasseSocial	Texto	Classe social do utilizador
InicioValidade	Data	Início da validade do registo
FimValidade	Data	Fim da validade do registo
Razao	Texto	Razão de fim da validade do registo
EmVigor	Número inteiro	Se o registo está em vigor (0) ou não (1)

Tabela B.4: Representação detalhada da dimensão Utilizador.

Campo	Tipo de Dados	Descrição
IdTipoMovimento	Número Inteiro	Identificador único de registo
Nome	Texto	Identifica o nome do movimento
Tipo	Texto	Identifica o tipo de transporte (aéreo, terrestre ou marítimo)
Natural	Número Inteiro	Identifica se o movimento é natural (1) ou não (0)
Motorizado	Número Inteiro	Identifica se o movimento é motorizado (1) ou não (0)
Coletivo	Número Inteiro	Identifica se o movimento é coletivo (1) ou não (0)

Tabela B.5: Representação detalhada da dimensão Tipo de Movimento.

Campo	Tipo de Dados	Descrição
IdTrajetoria	Número Inteiro	Identificador único de registo
CodigoTrajetoria	Número Inteiro	Identificador único da trajetória
DistanciaPercorrida	Número Decimal	Valor que representa a distância total da trajetória
VelocidadeMedia	Número Decimal	Valor que representa a velocidade média da trajetória
TempoDecorrido	Número Inteiro	Valor que representa o tempo total da trajetória

Tabela B.6: Representação detalhada da dimensão Trajetória.

Campo	Tipo de Dados	Descrição
IdDispositivoCaptura	Número Inteiro	Identificador único de registo
Tipo	Texto	Identifica o tipo de dispositivo (por exemplo, telemóvel)
Marca	Texto	Marca do dispositivo
Modelo	Texto	Modelo do dispositivo
Precisao	Texto	Precisão do dispositivo de captura

Tabela B.7: Representação detalhada da dimensão Dispositivo de Captura.

Campo	Tipo de Dados	Descrição
IdPontoEstadia	Número Inteiro	Identificador único de registo
Latitude	Número Decimal	Valor que representa a latitude do conjunto de coordenadas do ponto de estadia
Longitude	Número Decimal	Valor que representa a longitude do conjunto de coordenadas do ponto de estadia
Nome	Texto	Nome (se existir) da localização do ponto de estadia
MoradaCompleta	Texto	Morada completa do ponto de estadia
Rua	Texto	Rua do ponto de estadia
Numero	Número Inteiro	Número do ponto de estadia
CodigoPostal	Texto	Código-Postal do ponto de estadia
Bairro	Texto	Bairro do ponto de estadia
Freguesia	Texto	Freguesia do ponto de estadia
Cidade	Texto	Cidade do ponto de estadia
Regiao	Texto	Região do ponto de estadia
Pais	Texto	Pais do ponto de estadia
Continente	Texto	Continente do ponto de estadia
InicioValidade	Data	Início da validade do registo
FimValidade	Data	Fim da validade do registo
Razao	Texto	Razão de fim da validade do registo
EmVigor	Número inteiro	Se o registo está em vigor (0) ou não (1)

Tabela B.8: Representação detalhada da dimensão Ponto de Estadia.

Campo	Tipo de Dados	Descrição
IdCategoria	Número Inteiro	Identificador único de registo
Categoria	Texto	Identifica a categoria do ponto de estadia (por exemplo, restaurante)

Tabela B.9: Representação detalhada da dimensão Categoria Ponto de Estadia.

Campo	Tipo de Dados	Descrição
FK_PontoEstadia	Número Inteiro	Identificador da tabela Ponto de Estadia
FK_CategoriaPontoEstadia	Número Inteiro	Identifica da tabela Categoria Ponto de Estadia

Tabela B.10: Representação detalhada da dimensão Ponte Ponto de Estadia.

Campo	Tipo de Dados	Descrição
FK_Data	Número Inteiro	Identificador da tabela Data
FK_Tempo	Número Inteiro	Identificador da tabela Tempo
FK_Localizacao	Número Inteiro	Identificador da tabela Localização
FK_PontoEstadia	Número Inteiro	Identificador da tabela Ponto de Estadia
FK_Utilizador	Número Inteiro	Identificador da tabela Utilizador
FK_Trajetória	Número Inteiro	Identificador da tabela Trajetória
OrdemTrajetoria	Número Inteiro	Valor que representa a ordem do conjunto de coordenadas da trajetória
FK_Utilizador	Número Inteiro	Identificador da tabela Tipo de Movimento
FK_Trajetória	Número Inteiro	Identificador da tabela Dispositivo de Captura
Latitude	Número Decimal	Valor que representa a latitude do conjunto de coordenadas do registo
Longitude	Número Decimal	Valor que representa a longitude do conjunto de coordenadas do registo
Velocidade	Número Decimal	Valor que representa a velocidade instantânea em metros por segundo do conjunto de coordenadas do registo
Aceleração	Número Decimal	Valor que representa a aceleração em metros por segundo do conjunto de coordenadas do registo
DistanciaPercorrida	Número Decimal	Valor que representa a distância em metros do conjunto de coordenadas do registo
TempoDecorrido	Número Inteiro	Valor que representa o tempo em segundos do conjunto de coordenadas do registo

Tabela B.11: Representação detalhada da Tabela de Factos Movimento.

Apêndice C

Processo ETL

Este anexo contém as tabelas parciais que resultaram do processo de transformação do processo ETL.

```
1;hospital
2;parking
3;health
4;establishment
5;route
6;intersection
7;local_government_office
8;university
9;school
10;point_of_interest
11;museum
12;restaurant
```

Figura C.1: Tabela parcial da dimensão *Categoria Ponto de Estadia*.

```
1;12;4;2007;2007-04-12;quinta-feira, 12 de Abril de 2007;5;Quinta-Feira;True;False;1;False;False;;Abril;Primavera
2;13;4;2007;2007-04-13;sexta-feira, 13 de Abril de 2007;6;Sexta-Feira;True;False;1;False;False;;Abril;Primavera
3;14;4;2007;2007-04-14;sábado, 14 de Abril de 2007;7;Sabado;False;True;1;False;False;;Abril;Primavera
4;15;4;2007;2007-04-15;domingo, 15 de Abril de 2007;1;Domingo;False;True;1;False;False;;Abril;Primavera
5;16;4;2007;2007-04-16;segunda-feira, 16 de Abril de 2007;2;Segunda-Feira;True;False;2;False;False;;Abril;Primavera
6;17;4;2007;2007-04-17;terça-feira, 17 de Abril de 2007;3;Terca-Feira;True;False;2;False;False;;Abril;Primavera
7;18;4;2007;2007-04-18;quarta-feira, 18 de Abril de 2007;4;Quarta-Feira;True;False;2;False;False;;Abril;Primavera
8;19;4;2007;2007-04-19;quinta-feira, 19 de Abril de 2007;5;Quinta-Feira;True;False;2;False;False;;Abril;Primavera
9;20;4;2007;2007-04-20;sexta-feira, 20 de Abril de 2007;6;Sexta-Feira;True;False;2;False;False;;Abril;Primavera
10;21;4;2007;2007-04-21;sábado, 21 de Abril de 2007;7;Sabado;False;True;2;False;False;;Abril;Primavera
11;22;4;2007;2007-04-22;domingo, 22 de Abril de 2007;1;Domingo;False;True;2;False;False;;Abril;Primavera
12;23;4;2007;2007-04-23;segunda-feira, 23 de Abril de 2007;2;Segunda-Feira;True;False;2;False;False;;Abril;Primavera
```

Tabela C.1: Tabela parcial da dimensão *Data*.

```
0;Desconhecido;Desconhecido;;
```

Tabela C.2: Tabela da dimensão *Dispositvo*.

```
1;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
2;1;2;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
3;1;3;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
4;1;2;4;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
5;1;3;4;5;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
6;1;5;6;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
7;1;5;6;7;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
8;1;5;6;7;8;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
9;1;2;4;9;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
10;1;2;3;4;5;6;7;8;9;10;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
11;1;5;6;7;11;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
12;1;3;4;5;9;10;11;12;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
```

Tabela C.3: Tabela parcial da dimensão *Localização*.

```
1;4
2;5
3;4
3;10
3;25
4;4
4;7
5;4
6;5
7;5
8;5
```

Tabela C.4: Tabela parcial da dimensão *Ponte Ponto de Estadia*.

```
0;;;;;;;;;;;;;;;2007-04-12;1900-01-01;;True
1;39,973213000000000;116,330185000000000;Shuangyushu North Road Unit;4? Shuangyushu North Road, Haidian, Beijing, China, 100086;Shuangyushu North Road;
2;40,000282700000000;116,322865500000000;Qinghua West Rd;Qinghua West Road, Haidian, Beijing, China, 100084;Qinghua West Road;;100084;;Haidian;Beijing;
3;39,961062500000000;116,355126600000000;?????????;Xitucheng Road Side Road, Haidian, Beijing, China, 100876;Xitucheng Road Side Road;;100876;;Haidian;
4;39,987287000000000;116,303486000000000;Haidian Bureau Of Sports Society Sports Management Center;2? Yiheyuan Road, Haidian, Beijing, China, 100080;Yi
5;39,924668000000000;116,472734000000000;Shuiduizi East Road Unit;3? Shuiduizi East Road, Chaoyang, Beijing, China, 100026;Shuiduizi East Road;3?;10002
6;39,898965600000000;116,433447900000000;Donghuashi N St;Donghuashi North Street, Dongcheng, Beijing, China, 100062;Donghuashi North Street;;100062;;Do
7;39,965328700000000;116,308753500000000;Wanquanhe Road Side Rd;50? Wanquanhe Road, Haidian, Beijing, China;Wanquanhe Road;50?;;;Haidian;Beijing;Beijin
8;39,966653200000000;116,302174700000000;Wanliu East Rd;11? Wanliu East Road, Haidian, Beijing, China, 100089;Wanliu East Road;11?;100089;;Haidian;Beij
9;39,900323000000000;116,426524000000000;CPC Chongwenmen East Street Community Party Committee;19? Beijing Station West Street, Dongcheng, Beijing, Chi
10;39,985352800000000;116,297814900000000;Wanquanhe Road Side Rd;Wanquanhe Road Side Road, Haidian, Beijing, China, 100091;Wanquanhe Road Side Road;;10
11;39,978566000000000;116,301014800000000;Wanquanhe Road Side Rd;Wanquanhe Road Side Road, Haidian, Beijing, China;Wanquanhe Road Side Road;;;;Haidian;
12;40,002287400000000;116,274130100000000;Suzhou Street;5?-7? Suzhou Street, Haidian, Beijing, China, 100081;Suzhou Street;5?-7?;100080;;Haidian;Beijin
```

Tabela C.5: Tabela parcial da dimensão *Ponto de Estadia*.

```
1;0;0;0;00:00:00;Madrugada
2;0;0;1;00:00:01;Madrugada
3;0;0;2;00:00:02;Madrugada
4;0;0;3;00:00:03;Madrugada
5;0;0;4;00:00:04;Madrugada
6;0;0;5;00:00:05;Madrugada
7;0;0;6;00:00:06;Madrugada
8;0;0;7;00:00:07;Madrugada
9;0;0;8;00:00:08;Madrugada
10;0;0;9;00:00:09;Madrugada
11;0;0;10;00:00:10;Madrugada
12;0;0;11;00:00:11;Madrugada
```

Tabela C.6: Tabela parcial da dimensão *Tempo*.

```
0;Desconhecido;;;;
1;Andar;Terrestre;True;False;False
2;Bicicleta;Terrestre;True;False;False
3;Autocarro;Terrestre;False;True;True
4;Carro;Terrestre;False;True;False
5;Metro;Terrestre;False;True;True
6;Comboio;Terrestre;False;True;True
7;Avião;Aéreo;False;True;True
8;Barco;Maritimo;False;True;True
9;Correr;Terrestre;True;False;False
10;Mota;Terrestre;False;True;False
11;Taxi;Terrestre;False;True;True
```

Tabela C.7: Tabela da dimensão *Tipo de Movimento*.

```
1;1;14938,621492747507;2,2424163576361784;29888
2;2;1303,660057921742;0,96163716437685942;2227
3;3;18648,325351339176;2,9305745693342136;4800
4;4;1895,6605873291719;1,933295750698123;665
5;5;8598,8240697020665;1,0252919710383412;15916
6;6;1700,9866634073931;1,4742773837794989;530
7;7;3490,3734598038523;3,7448160978273797;965
8;8;424,49427700140313;2,4239192424205434;145
9;9;14947,294380280058;1,3130864094437025;14955
10;10;8288,0346899537999;4,1169003730713705;7775
11;11;11768,551013261473;1,4803506175042476;8330
12;12;11248,087532610456;2,2788151057549877;9551
```

Tabela C.8: Tabela parcial da dimensão *Trajetória*.

```
1;0;1;;;;;;;;;2007-04-12;;;True
2;1;2;;;;;;;;;2007-04-12;;;True
3;2;3;;;;;;;;;2007-04-12;;;True
4;3;1;;;;;;;;;2007-04-12;;;True
5;4;1;;;;;;;;;2007-04-12;;;True
6;5;3;;;;;;;;;2007-04-12;;;True
7;6;3;;;;;;;;;2007-04-12;;;True
8;7;3;;;;;;;;;2007-04-12;;;True
9;8;3;;;;;;;;;2007-04-12;;;True
10;9;2;;;;;;;;;2007-04-12;;;True
11;10;1;;;;;;;;;2007-04-12;;;True
12;11;3;;;;;;;;;2007-04-12;;;True
```

Tabela C.9: Tabela parcial da dimensão *Utilizador*.

Apêndice D

Validação do Modelo Proposto

Este anexo contém informação variada sobre o capitulo relativo à validade do modelo proposto nesta dissertação.

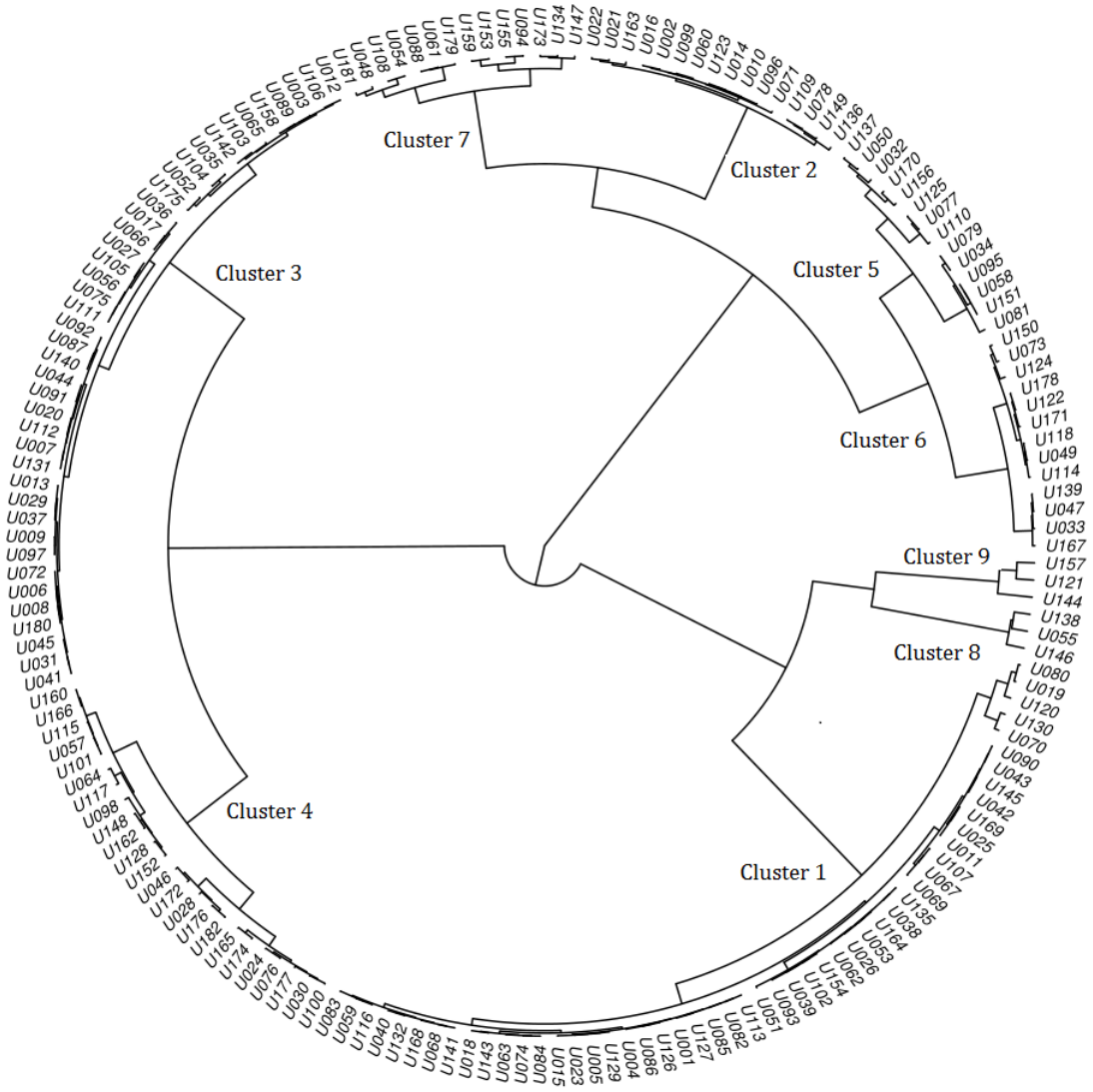


Figura D.1: Dendograma circular resultante da rotina de descoberta de utilizadores semelhantes.

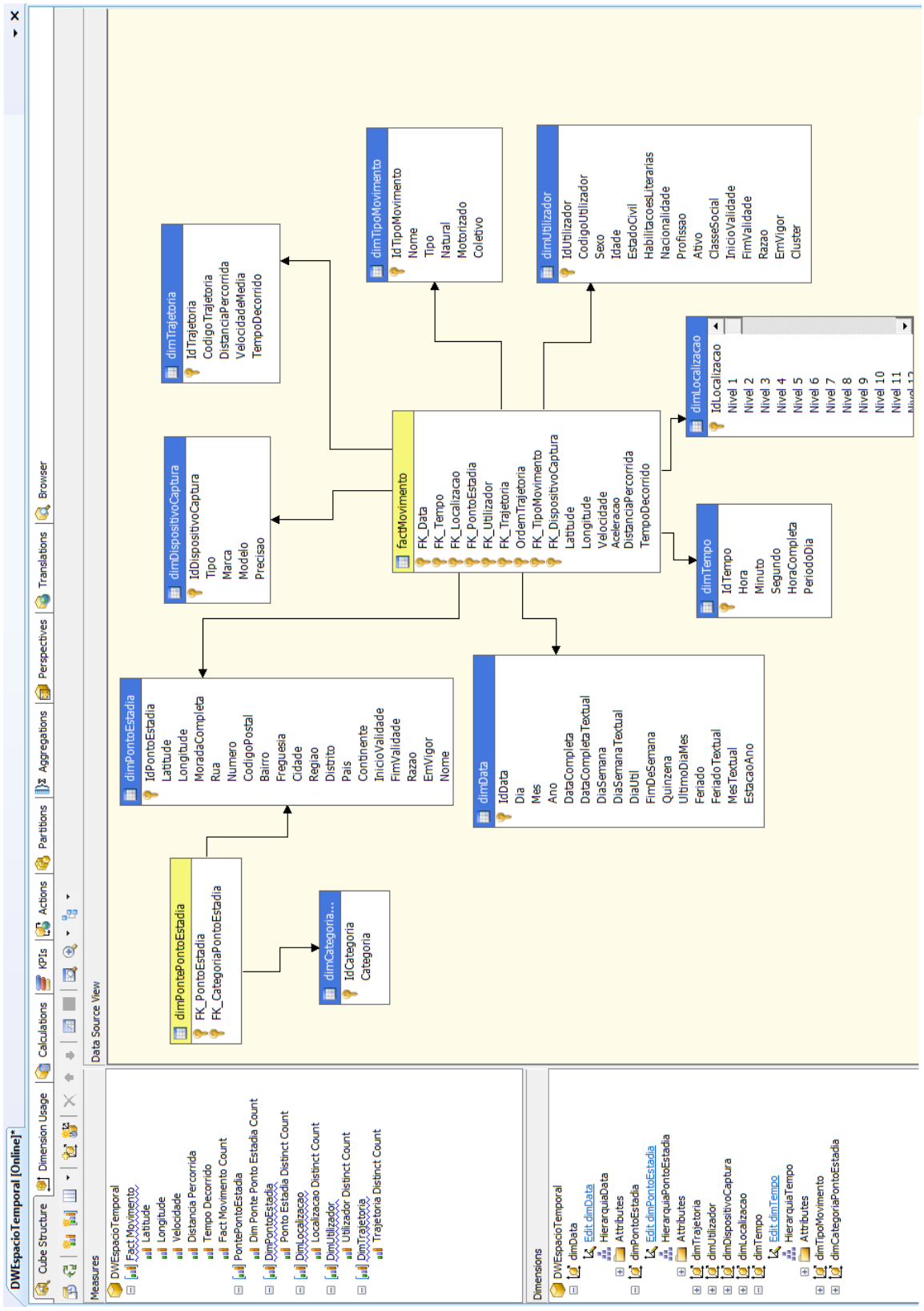


Figura D.2: Ambiente de desenvolvimento do cubo de dados.

Abreviaturas

CSV	*Comma-separated values*
DBSCAN	*Density-based spatial clustering of applications with noise*
DW	*Data Warehouse*
DWTrs	*Data Warehouse* de Trajetórias
EIS	*Executive Information Systems*
ETL	*Extract Transform Load*
GMT	*Greenwich Mean Time*
GPS	*Global Positioning System*
HOLAP	*Hybrid On Line Analytical Processing*
LaSige	*Large-Scale Informatics Systems Laboratory*
MB	*Megabyte*
MOLAP	*Multidimensional On Line Analytical Processing*
OLAP	*On-line Analytical Processing*
OLTP	*Online Transaction Processing*
PHP	*Hypertext Preprocessor*
PLT	*Track File*
ROLAP	*Relational On Line Analytical Processing*
SGBD	Sistema de Gestão e Base de Dados
SIG	Sistema de Informação Geográfica
SInteliGIS	*Services for Intelligent Geographical Information Systems*

Bibliografia

[1] elifelog.org - open space for life-log research collaboration. `https://www.elifelog.org/book/microsoft-geolife-gps-trajectories`. Acedido em 15-10-2012.

[2] LaSIGE - Large-Scale Informatics Systems Laboratory. `http://lasige.di.fc.ul.pt/`. Acedido em 20-06-2013.

[3] SInteliGIS - Services for Intelligent Geographical Information Systems. `http://xldb.fc.ul.pt/wiki/Sinteligis`. Acedido em 20-06-2013.

[4] Carlos Almeida. Data warehouse de trajetórias: um modelo semântico com suporte à agregação por direção dos movimentos. Master's thesis, Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brasil, 2010.

[5] Luis Alvares, Vania Bogorny, Bart Kuijpers, Jose de Macedo, Bart Moelans, and Alejandro Vaisman. A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th annual ACM International Symposium on Advances in Geographic Information Systems*, GIS '07, pages 1–8, New York, NY, USA, 2007.

[6] Richard Becker, Ramon Caceres, Karrie Hanson, Ji Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011.

[7] Pavel Berkhin. A survey of clustering data mining techniques. *Grouping Multidimensional Data*, pages 25–71, 2006.

[8] Sotiris Brakatsoulas. Modeling, storing and mining moving object databases. In *In Proc. IDEAS conference*, pages 68–77, 2004.

[9] Fernando Braz, Salvatore Orlando, Renzo Orsini, Alessandra Raffaeta, Alessandro Roncato, and Claudio Silvestri. Approximate aggregations in trajectory data warehouses. In *Proceedings of the IEEE 23rd International Conference on Data Engineering Workshop*, ICDEW '07, pages 536–545, Washington, USA, 2007.

[10] Edgar Codd, Sharon Codd, and C. Salley. Providing OLAP (On-Line Analytical Processing) to User-Analysis: An IT Mandate, 1993.

[11] Yves-Alexandre de Montjoye, César Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, March 2013.

[12] Michel Deza and Elena Deza. *Encyclopedia of Distances*. Springer, 1st edition, August 2009.

[13] Somayeh Dodge, Robert Weibel, and Anna-Katharina Lautenschütz. Towards a taxonomy of movement patterns. *Information Visualization*, 7(3):240–252, June 2008.

[14] Bo Fan. A hybrid spatial data clustering method for site selection: The data driven approach of gis mining. *Expert Syst. Appl.*, 36(2):3923–3936, 2009.

[15] Alberto Ferrari and Marco Russo. The many-to-many revolution advanced dimensional modeling with microsoft sql server analysis services. 2011.

[16] António Ferreira. Integração e processamento analítico de informação. Slides da disciplina, 2012.

[17] Fosca Giannotti and Dino Pedreschi. *Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[18] Mattia Gustarini and Wac Katarzyna. Estimating people perception of intimacy in daily life from context data collected with their mobile phone. *Pervasive 2012*, 2012.

[19] Jiawei Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[20] Jiawei Han, Micheline Kamber, and Anthony Tung. Spatial clustering methods in data mining: A survey. In Harvey J. Miller and Jiawei Han, editors, *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS*. Taylor and Francis, 2001.

[21] Francois Husson, Sebastien Lê, and Jerome Pagès. *Exploratory Multivariate Analysis by Example Using R*. A Chapman & Hall book. CRC Press, 2011.

[22] William Inmon. *Building the Data Warehouse*. John Wiley & Sons, Inc., New York, NY, USA, 1992.

[23] Christian Jensen, Augustas Kligys, Augustas Kligys, Torben Bach Pedersen, Torben Bach Pedersen, Igor Timko, and Igor Timko. Multidimensional data modeling for location-based services. *The VLDB Journal*, 13:1–21, 2002.

[24] Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 9th edition, March 1990.

[25] Ralph Kimball and Joe Caserta. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley, Indianapolis, IN, 2004.

[26] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons, Inc., New York, NY, USA, 2nd edition, 2002.

[27] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, GIS '08, pages 34:1–34:10, New York, NY, USA, 2008.

[28] Jamie MacLennan, ZhaoHui Tang, and Bogdan Crivat. *Data Mining with Microsoft SQL Server 2008*. Wiley Publishing, 2008.

[29] Gerasimos Marketos, Elias Frentzos, Irene Ntoutsi, Nikos Pelekis, Alessandra Raffaetà, and Yannis Theodoridis. Building real-world trajectory warehouses. In *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, MobiDE '08, pages 8–15, New York, NY, USA, 2008.

[30] Vitor Oliveira, Ana Paula Afonso, and André Falcão. Conceção de um data warehouse espácio-temporal para análise de trajetórias humanas. In *Proceedings do quinto Simpósio Português de Informática*, InForum '13, Évora, Portugal, 2013.

[31] Salvatore Orlando, Renzo Orsini, Alessandra Raffaetà, Alessandro Roncato, and Claudio Silvestri. Trajectory data warehouses: Design and implementation issues. *Journal of Computing Science and Engineering*, 1(2):211–232, 2007.

[32] Carlo Ratti, Riccardo Pulselli, Sarah Williams, and Dennis Frenchman. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

[33] Ahmad Raza, Shafqat Hameed, and Tim Macintyre. Global positioning system – working and its applications. In Khaled Elleithy, editor, *Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering*, pages 448–453. Springer Netherlands, 2008.

[34] Kai-Florian Richter, Falko Schmid, and Patrick Laube. Semantic trajectory compression: Representing urban movement in a nutshell. *J. Spatial Information Science*, 4(1):3–30, 2012.

[35] Salvatore Rinzivillo, Dino Pedreschi, Mirco Nanni, Fosca Giannotti, Natalia Andrienko, and Gennady Andrienko. Visually driven analysis of movement data by progressive clustering. *Information Visualization*, 7(3):225–239, June 2008.

[36] Maria Silva. Tetl: Uma ferramenta de etl para trajetórias de objetos móveis. Master's thesis, Universidade Federal de Pernambuco, Recife, Brasil, 2009.

[37] Roger Sinnott. Virtues of the Haversine. *Sky and telescope*, 68:158, 1984.

[38] Stefano Spaccapietra, Christine Parent, Maria Luisa Damiani, Jose Antonio de Macedo, Fabio Porto, and Christelle Vangenot. A conceptual view on trajectories. *Data Knowledge Engineering*, 65(1):126–146, April 2008.

[39] Joe Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[40] Chris Webb, Alberto Ferrari, and Marco Russo. *Expert Cube Development with Microsoft SQL Server 2008 Analysis Services*. Packt Publishing, 2009.

[41] Yu Zheng. *User Guide Geolife Dataset*. Microsoft Research Asia, version 1.3 edition, 8 2012.

[42] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, UbiComp '08, pages 312–321, New York, NY, USA, 2008.

[43] Yu Zheng and Xing Xie. Learning Location Correlation from GPS Trajectories. In *MDM '10: Proceedings of the 2010 Eleventh International Conference on Mobile Data Management*, pages 27–32, Washington, DC, USA, 2010.

[44] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 791–800, New York, NY, USA, 2009.

[45] Yu Zheng and Xiaofang Zhou. *Computing with Spatial Trajectories*. Springer, 2011.